

CHAPTER 2

DATA AND DATABASES

Fumiaki Katagiri and Adam Arkin

INTRODUCTION

Data does not appear to be in short supply in contemporary biology. The development of high-throughput technologies, in particular, has generated massive amounts of information. While these technologies produce information about the chemistry of a system, as with the sequence and structure databases, the biological status of the organism is often of little importance. However, when the goal is describing a process, such as signal transduction or gene expression, the information gathered will be highly conditioned by cell type, experimental conditions, and other variables. Since the modeling of dynamic biological processes is a central aspect of systems biology, the nature of the data available to create and test models is of great importance. This chapter will examine some aspects of data generation and data storage and access, as it applies to systems biology. It will not focus on the technologies themselves. Systems biology relies on a wide variety of data types. Some of these have become fairly standard, such as sequence and structure information, mass spectrometry, and microarrays. Others are still in the process of development, such as single cell measurements using multicolored fluorescent assays, fluorescent antibodies and covalent conjugates, and quantum dots using high-resolution microscopy and flow cytometry. The diversity of tools used precludes a detailed discussion in this chapter. Additionally, although broad access of all tools to investigators is an issue, there do not appear to be systematic differences between countries in the availability of technologies.

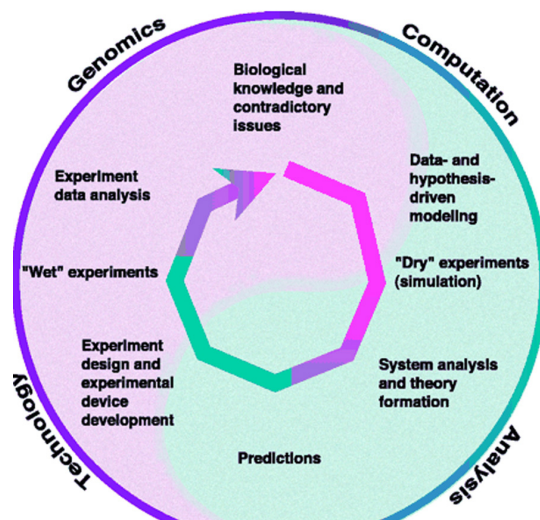


Figure 2.1. Hypothesis-driven research in systems biology (Kitano, 2002).

The interaction of experiment and models requires accurate datasets to infer network structures, to create the models, and to test and distinguish the predictions from multiple models. Successive iterations of model building, prediction, experiment, and subsequent refinements of the models are the result. This repeating cycle of experimentation and theoretical work is the engine that pushes the progress of systems biology research (Kitano, 2002). Consequently, while the concept of systems biology is not all that new, one reason

that it has displayed renewed vigor is the impressive advance of experimental technologies. The development and spread of various approaches to high-throughput measurements have been contributing to generation of a large amount of systematic data from a wide variety of biological systems. These data have fueled high-throughput discovery research based on reductionist approaches. Furthermore, rapid generation of such data has given people a sense of hope that we may now be able to collect sufficient information to understand biological phenomena as behaviors of dynamic systems. Is this hope built on a solid foundation?

Roughly speaking, we can distinguish two stages in systems biology research. The type of experimentations useful in each stage is also distinct. When the network structure of interest is not well-defined, systematic and broad-spectrum characterizations, such as global profiling, provide the necessary information. When the network structure is well defined and quantitative models are built based on the network structure, the experimental information demanded is very specific to the network of interest, the proposed models, and the questions asked about the network. In such a case, the data collection is model-driven and experiments need to be designed according to the specific demands. In addition, the techniques used must, in most cases, provide quantitative results.

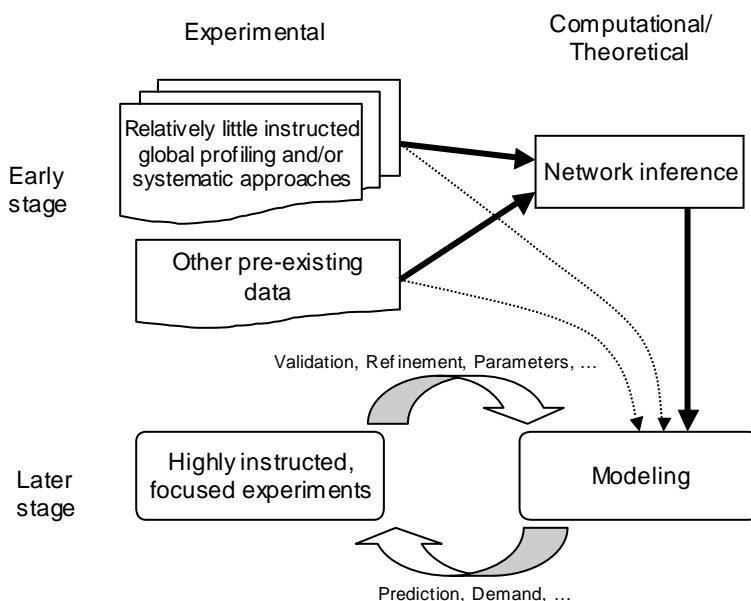


Figure 2.2. The two different stages in systems biology study and the types of useful experimental approaches for each stage.

In both types of experimentation, it is crucial to use a combination of methods to perturb the network and to measure the effects of the perturbations. When no model is available, there is little *a priori* information about the network to limit the set of targets for measurement. Consequently, the ideal is to densely cover the space of all the possibilities in both perturbation and measurement. The ideal perturbation method thus enables perturbation of every node or edge in the network specifically and quantitatively, and the ideal measurement methods measure all the parameters that define the state of the network. Practically, systematic genetic perturbation is often used as a perturbation method. Ideally, for the initial phase of determining overall regulatory organization and inferring what should be in the model, four criteria are important. First, the data should be exhaustive within each category of measurements, such as the measurement of messenger ribonucleic acid (mRNA) levels, measurement of protein levels, measurement of cell sizes, etc. For example, when mRNA levels are measured, the mRNA levels of all the genes in the organism should be measured. If the measurement must be limited to a subset of the genes, we should have a good idea that the subset likely contains all the genes that are important for the network of interest. Second, many different categories, ideally all the possible categories, of data should be collected in a correlated manner. Correlative measurements in different categories are crucial for integration of various measured events into a single network because such data are context-dependent. For example, if a protein-protein interaction measurement is performed under one condition, the measurement may not be useful under a different condition. Third, the data should have sufficient resolution in time and space. If a measurement does not have sufficiently dense

time points, information about the dynamics of the system is limited. Considering an example of spatial resolution, if a measurement is performed with a mixture of different cell types that behave differently, the resulting averaged measurement fails to detect distinct cell-type modules. Fourth, the measurement should be quantitative. Whereas binary Boolean models have their own utility in many cases, they are often not sufficient to capture important network dynamics.

One major trend that is initiated by genomics research is the development of various highly parallel measurement (broad profiling) technologies. RNA profiling methods allow reasonably complete measurements in many organisms in which the genome sequences are known. The use of microarrays and reverse transcription-polymerase chain reaction (RT-PCR) results in good sensitivity and accuracy. Protein and metabolite profiling methods are improving quickly by combining chromatographic or electrophoretic separation methods with mass spectrometry-based methods. Although our study did not cover them, microfluidics and other micro-manufacturing technologies are expected to dramatically improve the cost, speed, and labor-intensiveness of highly parallel measurements. However, we are still missing good profiling methods for many categories of data. For example, if we want to know the amount of a particular modified form of a particular protein in a particular subcellular location, we still need to perform focused research. And these data types are often the most useful for mechanistic modeling of pathways. Another issue with broad profiling technologies is that they do not provide sufficient accuracy in many cases. In addition to technical challenges, practical issues, such as limitations in budget, time, and human resources, could limit the kinds of data obtainable by broad profiling methods.

The discussion in this chapter will be on both large-scale data collection and profiling technologies as well as smaller, model-focused data. However, the more global databases, which are largely independent of the biological state of the systems, clearly predominate. It must be noted that there are remarkably few examples of the melding of modeling and measurement technology in the U.S. or Japan, and the large-scale databases are rarely useful for systems biology studies. In Europe, this melding is an essential feature of EU-funded research consortium projects (COMBIO, COSBICS, DIAMONDS, EMI-CD). A new initiative has been started (ENFIN) to apply the large databases to systems biology research. Incompleteness of information, e.g., limitations in quantitation, accuracy, resolution, and the categories of data, is the major reason that data generated by broad profiling technologies are underutilized in systems biology research. A large number of data points with limited measurement accuracy also present challenges in having sufficient statistical power in analysis. Applications of detailed systems biology modeling are so far often limited to relatively small, well-isolated networks in which it is practical to perform targeted, intensive experimentation to obtain key information to answer specific questions.

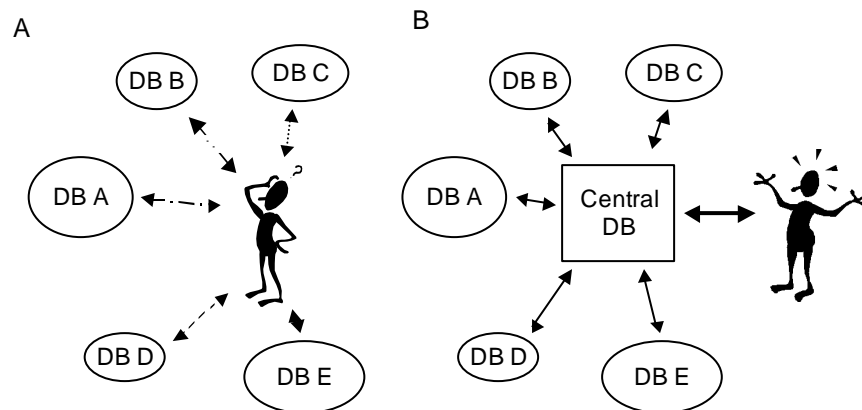


Figure 2.3. The importance of standardization and virtual consolidation of databases.

Exhaustive, quantitative measurements with high resolution in many categories necessarily generate a large amount of data. To effectively integrate such data into models of biological systems, powerful database platforms are essential. Archiving quickly growing, context-dependent data in a usable manner is a serious challenge. Curating a large amount of diverse information for many different systems requires a wide range of knowledge. These situations tend to drive the generation of many, relatively small, highly focused databases. However, having many small databases presents a funding challenge for maintaining and updating

such databases as well as creating a standardization issue. Standardization is crucial in not only maintaining the integrity of the entire body of related databases but also in making them usable to general users. It is also an issue in the peer review of results and methods.

We will compare activities in producing and archiving data of importance to systems biology in the U.S., Europe, and Japan. This includes not only scientific aspects but also social background and funding situations that may have had significant influence in shaping the research activities in these regions. We will also consider general database issues apart from regional considerations, since the main issue is standardization at the global level. In the end, we will point out specific needs and recommend some possible directions for the experimental part of systems biology research.

RESEARCH ACTIVITIES IN THE U.S.

Databases

Large-scale technologies began with the sequencing programs directed to the understanding of the human genome. Following on their successes there emerged a variety of genome-wide measurements of cellular function. These include whole genome gene expression microarrays, large-scale deletion libraries, structure, and the application of mass spectrometry to assess, for example, protein-protein interactions. These in turn have led to the creation of databases containing a variety of data types, including organism-specific databases supported by federal funding, usually from the National Institutes of Health (NIH), such as *Saccharomyces cerevisiae* (<http://www.yeastgenome.org/>); WormBase (<http://www.wormbase.org/>) for *C.elegans* and other nematodes; FlyBase (<http://flybase.bio.indiana.edu/>) for *Drosophila melanogaster*; and The Arabidopsis Information Resource (<http://arabidopsis.org/>) for *Arabidopsis thaliana*. These databases include a diverse array of data types, such as fitness of deletion mutants (the yeast database) and anatomy and spatial expression data (FlyBase and WormBase). They represent significant ongoing investments in the United States in centralizing and standardizing data necessary for systems biology. The co-evolving efforts associated with this are the somewhat spotty and erratic investments in data standards such as the microarray data standards outlined at the Microarray Gene Expression Data Society (MGED) (<http://www.mged.org/>); ontologies for describing gene function or structure with a controlled vocabulary (<http://www.geneontology.org/>); and pathway information (<http://www.biopax.org/>) and biological model (<http://sbml.org/index.psp>) storage and transport.

At the next level from these comprehensive databases are national investments in smaller, more focused, but still relatively large-scale projects. These include support from the National Institute of General Medical Sciences of the NIH for so-called “Glue Grants” which require coordination of multi-site activities around specific targeted goals, in many of which systems biology plays a role. These include the Cell Motility Consortium (<http://www.cellmigration.org/>) dedicated to developing reagents, measurements, and models of cell migration; the lipid MAPS Consortium (<http://www.lipidmaps.org/>), to identify, quantitate, and define the interactions of cellular lipids; the Consortium for Functional Glycomics (www.functionalglycomics.org/), to understand the role of carbohydrates in cell-cell communication; and the Alliance for Cellular Signaling (<http://www.signaling-gateway.org/>), to examine signal transduction pathways. Other such tightly focused efforts include the Alpha Project (<http://www.molsci.org/>), supported by the National Human Genome Research Institute, to look at the pheromone response in yeast and the Department of Energy’s Genome to Life (GTL) projects that seek to understand the mechanisms by which microbes function in the environment.

Data Analysis

Much of the technology that was at one time thought to be too expensive for individual investigators (and thus needed to be centralized) has, in fact, found its way into many laboratories. The most prevalent of these are microarrayers, but there are also improved fluorescent microscopes, small flow cytometers, and even mass-spectrometers that have come down in price far enough such that a single lab can own one. In many cases, the experiments for any particular system need to be optimized such that even a university central facility would be inappropriate (unlike sequencing, which can be easily outsourced). Thus, the rate of data generation (and the diversity of experimental protocols and reporting formats) has exploded and driven a large boom in the academic (and industrial) data analysis efforts. These include the development of statistical methods and experimental designs specialized in ferreting out sources of systematic and random error in microarray experiments, typified by the statistically grounded methods of Trevor Hastie and Robert

Tibshirani at Stanford (Hastie and Tibshirani, 2004), and Terry Speed at the University of California Berkeley (Bolstad et al., 2003). It extends to sophisticated methods for correlation and fusion of data across experimental conditions and different measurement types to derive the “modules” of co-regulation and their relationships, typified by the developments in statistical graph modeling approaches championed by Daphne Koller (Stanford) (Segal et al., 2003), David Gifford (MIT) (Bar-Joseph et al., 2003), and Michael Jordan (U.C. Berkeley) (Lanckriet et al., 2004). With these analytical methods as a foundation, together with data on the upstream sequence of co-expressed genes, and knowledge of protein structural domain interactions, these correlative methods are being applied to the network inference problems described in the third chapter. Additionally, John Doyle at Caltech (Carlson and Doyle, 2002), Frank Doyle at University of California, Santa Barbara (UCSB) (Gunawan et al., 2005), and Michael Frenklach (Rao et al., 2004) and Andrew Packard at U.C. Berkeley are all beginning to employ statistically and physically grounded data analyses for full dynamic model parameterization, model validation and model discrimination. These methods are expected to become more important as more quantitative data and models are developed and need to be formally compared.

Modeling and Data Collection

The Alpha Project and the Alliance for Cell Signaling represent two attempts to systematically collect data in well-defined systems together with efforts to create models to predict system behavior. There are a relatively small number of laboratories that are engaged in developing models and concomitantly testing them experimentally. The systems studied are quite diverse, ranging from bacterial chemotaxis (Alon et al., 1999) to the epidermal growth factor (EGF) receptor in fibroblasts (Wiley et al., 2003) to the Wnt signaling pathway in *Xenopus* (Lee et al., 2003). Since each of these represent a different biological system, a centralized database is not easy to justify for this kind of data collection. However, this does not mean that standardization and easy access to the data and computational methods is not required. Indeed, it is essential to effectively review manuscripts and develop statistics and algorithms to analyze and cross-compare experiments from a variety of laboratories. In whatever form, databases need to be established that conform to common standards of data collection and ontology, and that contain enough meta-data defining conditions of measurement to allow such evaluation and comparisons. This will be discussed in more detail later.

RESEARCH ACTIVITIES IN EUROPE

The panel saw a variety of styles of organizations in data and database aspects of systems biology research in Europe in July 2004. They include large institutes that are directed toward common goals, large consortiums, and small groups. The driving forces of different organizations are also often different. In some cases, the leadership of particular individuals was the key. In other cases, funding initiatives were the major factors.

Databases

Large-scale databases of sequence and structure are as common in Europe and the U.K. as in the U.S. One example is the European Bioinformatics Institute (EBI) located in Cambridge (<http://www.ebi.ac.uk/>). The Institute’s focus used to be database technologies that handle and facilitate use of a large amount of data generated by genomics research. The focus has been shifting toward studies aimed at biological functions while taking advantage of the Institute’s strength in computer technology. The Institute now includes research groups conducting biological experimentation. This shift from a purely computational institute to an institute with capabilities in both computation and experimentation reflects the recognition by computer scientists of the importance of close integration between experimental and theoretical work.

A large, focused operation is generally best conducted at a single site. Typically, this type of operation is led by a strong leader with a clear vision of the goals to be achieved. For example, the Max Planck Institute for Molecular Plant Physiology (<http://www.mpimp-golm.mpg.de/>) is focused on collecting correlated data, such as expression and metabolite profiles, from a large collection of genetically perturbed Arabidopsis plants. Usually, each department in a single Max Planck Institute operates independently. However, with the strong leadership of Dr. Willmitzer, this Max Planck Institute has been shaped toward a common goal of understanding plant metabolism. Two major departments led by Drs. Willmitzer and Stitt, together with other departments, cooperatively conduct these large data generation and analysis operations.

A large consortium can also be organized with geographically dispersed laboratories. The Hepatocyte Alliance in Germany was driven by the Federal Ministry of Education and Research (BMBF) funding

initiative “Systems for Life-Systems Biology.” The alliance has 25 total participating groups. Funding of €14 million is provided over the three years beginning in 2004. Standardization of biological materials within the alliance was rigorously implemented. The alliance works on a number of sub-projects, including detoxification/dedifferentiation and regeneration. In the regeneration sub-project, experiments in different groups are organized according to different signaling pathways. The approach of segmenting the project into smaller, defined subnetworks allows the groups working on different projects to be moderately independent while still effectively interactive despite geographic separation among the consortium members. The effectiveness of such an operation in coordinating the production and maintaining the quality of data is still unclear, since it was only a few months underway at the time of our visit. The benefits of this approach are the inclusion of a number of high-quality groups with specialized expertise that may not be available at a single site. However, dividing this type of work among multiple sites makes quality control of resources and data more difficult and increases the chances of mistakes/accidents in tracking them. The overall cost- and time-efficiency in production is also lower.

Modeling and Data Collection

As in the U.S., there are relatively few groups with closely linked efforts in modeling and experimentation where we could identify systematic data collection efforts. However, one prominent example is the group headed by Denis Noble at Oxford. They have a long history of studying the heart using experimental and modeling approaches at multiple scales from molecular to physiological aspects. While his laboratory has generated a large amount of data, specific demands of the model often call for different expertise. Therefore, various collaborations were developed at different stages of the research (<http://www.physiome.org/>), notably including the anatomic studies of Peter Hunter in New Zealand (<http://www.bioeng.auckland.ac.nz/home/home.php>). The data collections have been specifically linked to the needs of the model.

RESEARCH ACTIVITIES IN JAPAN

Funding for basic research by the Japanese government has dramatically increased in the last decade. Research in genomics, related high-throughput discovery research, and the bioinformatics supporting them has benefited from the increased financial support. Japan is particularly strong in the development of large databases.

Databases

The panel visited several large research institutes with strong database components during our visit in December 2004. They include the RIKEN Yokohama Institute (<http://www.yokohama.riken.jp/indexE.html>), which represents an effort under the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). Among the databases supported by RIKEN is the Mouse Genome Encyclopedia (<http://genome.gsc.riken.go.jp/>), which collects the sequences, physical clones, gene expression profiles, and protein-interaction data generated by members of the FANTOM (Functional Annotation of Mouse complementary deoxyribonucleic acid (cDNA)) consortium. RIKEN also has the Arabidopsis Genome Encyclopedia (<http://rarge.gsc.riken.go.jp/>), which includes collections of mutants and full-length cDNA clones, microarrays, and shape phenotypes of the mutants. In the next phase the Arabidopsis project will integrate genome to phenome and metabolome. In 2004 RIKEN initiated the Genome Network Project (<http://www.mext-life.jp/genome/english/index.html>) to identify transcriptional regulatory networks in the human genome. This is a national project, where the biological projects and technology development is performed by 12 groups of independent investigators while RIKEN provides the core data and resource production capability.

Dr. Shibata’s group at the Kazusa DNA Institute, which is mainly funded by Chiba prefecture, leads a collaborative effort to investigate metabolic networks in plants (Hirai et al., 2005). This project is also supported by NEDO (New Energy and Industrial Technology Development Organization), which is operated by the Ministry of Economy, Trade, and Industry (METI) and focuses on Arabidopsis. A particularly interesting database developed by Dr. Kanaya (Nara Institute of Science and Technology) is directed to candidate compound identification, and includes 20,000 microbial compounds and 100,000 plant compounds, and contains a variety of compound information as well as the species in which a particular compound has been detected.

The Japan Biological Information Research Center (JBIRC) is part of the National Institutes of Advanced Industrial Science and Technology (AIST), which operates under METI. It includes an integrated human genome annotation database (H-Invitational DB, <http://www.h-invitational.jp/>), containing information on 41,118 full-length cDNA clones including gene structures, functions, domains, expression (in some cases), diversity, and evolution.

Finally, mention must be made of the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (<http://www.genome.jp/kegg/>) of metabolic networks developed and maintained by Dr. Kanehisa's group at Kyoto University, which is widely used throughout the world.

The infrastructure to support these efforts is uniformly outstanding and frequently astonishingly good.

MODELING AND DATA COLLECTION

Most of the data development efforts come from high-throughput discovery research and technology development, and, once again, the number of groups that focus on the interaction of models and experiments in systems biology is still small in Japan as elsewhere. However, there are a number of examples of this kind of activity. Under the strong leadership of Dr. Kodama, the Laboratory of Systems Biology and Medicine at the University of Tokyo (http://www.lsbm.org/site_e/univ/) is streamlined for discovery and production of diagnostic and therapeutic antibodies. As part of this Dr. Ihara's group developed a protein interaction database through text mining to help understand mechanisms and select protein targets against which to raise antibodies. The Symbiotic Systems Project (<http://www.symbio.jst.go.jp/symbio2/index.html>) headed by Dr. Kitano has a number of programs that closely tie experiment and models. He is also working with a consortium that involves developing a database of automated recording of cell lineage in *C. elegans* mutants. The Institute for Advanced Biosciences (IAB, <http://www.iab.keio.ac.jp/>) at Keio University in Tsuruoka was built to fulfill Dr. Tomita's vision of the experimental needs of systems biology, so large-scale experimentation in the Institute is closely connected with theoretical work in the study of *E. coli* metabolism. In this project, Dr. Mori's collection of systematic mutants of *E. coli* is an outstanding resource (Mori, 2004). Dr. Ueda at the RIKEN Center for Developmental Biology (CDB) is a young PI building a research program for study of circadian rhythms in mammalian cells that involves high-throughput measurements and theoretical work (Ueda et al., 2005).

Many of our Japanese hosts acknowledged that much of their research does not fit our definition of systems biology. However, this situation may be changing. Dr. Yao, who is a consultant for the RIKEN Yokohama Institute, JBIRC, and CBRC, reported in our final workshop the MEXT plan for support focused on systems biology research. MEXT clearly considers systems biology as a next step after the establishment of high-throughput research infrastructure, which has been heavily funded in the past decade. If the program is well implemented, especially by facilitating involvement of more modeling-type researchers, Japan, with a high level of infrastructure, has great potential to make rapid progress in systems biology research.

Conclusions

There are two opposite trends in database organizations: large inclusive databases and small specialty databases. Both approaches have advantages and disadvantages. Large-scale databases, which primarily collect information that is not closely tied to the state of a cell, have become quite common, and their value is well understood (How useful these data are for systems biology is less clear. In general, the degree of quantitation is too limited to be used by investigators developing and testing models of biological processes). Standardization, although not complete, is progressing. This is not the case for many other kinds of data, particularly those tied to biological processes that are strongly conditioned by the state of the cell. It is not even clear how much "meta-data" is needed. Gene expression, protein expression, molecular localization, interactions, and post-translational modification are highly conditional. Indeed, the strain of cell used, the media, and other measurement conditions can appreciably affect the measured outcomes. There are a number of related issues, such as the amount of raw data needed, and the availability of statistical analysis and software packages used. These issues are not unique to data used for systems biology, but their absence is even more critical than in the analysis of state-independent data.

As noted in the introduction to this chapter, the variety of biological systems used for modeling in systems biology tends to drive the creation of many small, highly focused databases. The size of a database and the level of manual curation are inter-dependent. One researcher cannot be an expert in many different biological

systems, and thus intensively and manually curated databases tend to be small, with good quality control of the data. However, small databases are often independently developed and have substantial overlaps with other databases, and they are not well standardized. The absence of standardization severely limits the utility of these databases. For researchers who are not very familiar with a particular biological system, it may not be very easy to find the most appropriate database for their purposes. Furthermore, it is difficult for a small database project to be continuously funded for maintenance and updates.

The major reason that manual curation is valuable in databases is that some types of data are not easily formatted according to fixed rules. This is evident when a primary source of data is not designed for transferring the data to databases. Descriptive data from literature is a typical example. Some efforts must be made to deal with this issue. First, multiple different terms can be used to describe the same thing, or, even worse, the same terms can be used to describe different things, depending on the context. Efforts to impose a controlled vocabulary, i.e. ontology projects, were initiated to make data stored in databases self-consistent. If a controlled vocabulary is imposed at the stage of generating primary data sources (e.g. literature), the difficulties of manual curation in collecting data will be eased. Second, the relationships among terms and the context in which terms are used are very difficult to automatically capture while they are crucial in collecting accurate data from literature. One option would be to have authors of a paper submit a formatted, database-friendly summary of the work at the submission of the manuscript. Third, the data needs to be in a form that will allow critical evaluation of its reliability. This function of expert curation is becoming more important as more data are becoming available which are not carefully quality-controlled.

For convenience to users of databases, it is desirable to have small databases virtually consolidated in a large database framework using the same standards. In other words, accessing multiple databases through a central database should seem almost seamless to users. In this way, it could be possible to maintain small databases after termination of their funding although updating them could still remain a challenge. A high level of integration could be difficult with already highly developed databases due to the difference in underlying database schemes. However, if we set general standards, it could be achieved with newer, small databases. It would be helpful if such hierarchical relationships of standardized databases were organized in a research community for each particular biological system and research field.

In summary, progress in systems biology requires a balance between bottom-up efforts, which are based on creativity and competition/collaboration of individual researchers and/or laboratories, and top-down organization efforts, which are necessary for standardization and efficient use of limited resources. Generally speaking, the current success of U.S. research is largely owed to emphasis on bottom-up efforts. However, in many aspects of experimental/data technology for systems biology, we need to develop more centralized resources.

The importance of correlating multiple kinds of data favors sample preparation performed at a single facility. Ideally, single identifiable samples should be used for measurements in many different categories. In addition, the same person at the same facility should perform a single category of measurement with different samples. These criteria lead to two organizational models for experimentation. In one model, sample preparation and all the measurements are to be performed at a single large experimentation center. The Max Planck Institute for Molecular Plant Physiology in Germany, RIKEN Yokohama Institute in Japan, and the Institute for Advanced Biosciences in Japan are examples of such large centers. The other model is a consortium of several facilities, each of which is exclusively specialized, e.g., one site for all the sample preparation and other sites for each category of measurement. This second example is not common and the panel saw few examples, except perhaps for the Hepatocyte Project in Germany and the Alliance for Cell Signaling in the U.S. The large center model has advantages in better communication among the involved members, a lower chance of mistake/accident in experiment/data tracking, and higher cost-efficiency in operation. The large center model can also offer an opportunity for better communication between researchers in experimental and theoretical work by having such people at the same site. As emphasized in the beginning of the chapter, close interactions between experimental and theoretical work is crucial in success of systems biology research. The large center model has a disadvantage in requiring larger initial investment and with less flexibility as an organization in the long term. It can, however, accommodate focused research efforts by having programs for *ad hoc* experimentation teams. The consortium model can do this easily by adding appropriate laboratories as consortium members. In either model, operations at each site need to be tightly controlled for high-quality data generation. This is where top-down organizations work

better. Such a center or consortium would be a major data generation site for a biological system, and, therefore, it is reasonable for it to take a lead in organizing “the” database for the particular biological system.

Even if we do not choose as extreme an option as a large center or consortium, top-down organizational efforts will become more important in data generation and management because of the need for standardization and easy access to investigators. This will impact research communities from a social viewpoint. In top-down organizations, the role of an individual becomes more team-oriented, and it will be more difficult to single out accomplishments made by the individual. This situation is not very compatible with current academic evaluation criteria for merit. We will need to establish a different set of criteria or a different career path, so that team-oriented researchers can develop their careers. In top-down organizations, leaders also need to have strong management skills, which are usually not taught during typical scientific training. Furthermore, to make the situation fair to researchers not involved in the center/consortium, rapid dissemination of data generated by the center/consortium should be enforced. Although rapid data dissemination in genome sequencing projects has been the norm, rapid data dissemination in large experiments that involve sophistication in designing and performing experiments is not yet very common. To ensure that this and other functions are optimally developed for the benefit of science, research communities will need to be involved in governance.

The future of data and database aspects in systems biology research lies more on cooperation than competition. This is necessary to effectively utilize limited funding and human resources. Cooperation among funding agencies at inter-program, inter-agency, and international levels will also be important to facilitate cooperation among researchers. The spirit of cooperation should be extended beyond academia and governments to industry. The panel saw much more involvement of industry in systems biology research in Japan and Europe (particularly in Japan) than in the U.S. The impression the panel got was that legal situations around intellectual properties may be different in Japan and Europe. Although the panel did not have a chance to closely study such legal issues, they are crucial in increasing the involvement of industry in the U.S.

REFERENCES

- Alon, U., Surette, M. G., Barkai, N., and Leibler, S. 1999. Robustness in bacterial chemotaxis. *Nature* 397: 168–171.
- Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. 2003. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 21: 1337–1342.
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193.
- Carlson, J. M., and Doyle, J. 2002. Complexity and robustness. *Proc Natl Acad Sci U S A* 99 Suppl 1: 2538–2545.
- Gunawan, R., Cao, Y., Petzold, L., and Doyle, F. J., 3rd. 2005. Sensitivity analysis of discrete stochastic systems. *Biophys J* 88: 2530–2540.
- Hastie, T., and Tibshirani, R. 2004. Efficient quadratic regularization for expression arrays. *Biostatistics* 5: 329–340.
- Hirai, M. Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D. B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., *et al.* 2005. Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *J Biol Chem* 280: 25590–25595.
- Kitano, H. 2002. Systems biology: a brief overview. *Science* 295: 1662–1664.
- Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. 2004. A statistical framework for genomic data fusion. *Bioinformatics* 20: 2626–2635.
- Lee, E., Salic, A., Kruger, R., Heinrich, R., and Kirschner, M. W. 2003. The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol* 1: E10.
- Mori, H. 2004. From the sequence to cell modeling: comprehensive functional genomics in Escherichia coli. *J Biochem Mol Biol* 37: 83–92.
- Rao, C. V., Frenklach, M., and Arkin, A. P. 2004. An allosteric model for transmembrane signaling in bacterial chemotaxis. *J Mol Biol* 343: 291–303.

- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34: 166–176.
- Ueda, H. R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y., Iino, M., and Hashimoto, S. 2005. System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat Genet* 37: 187–192.
- Wiley, H. S., Shvartsman, S. Y., and Lauffenburger, D. A. 2003. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol* 13: 43–50.