Seung Yon Rhee
Carnegie Institution of Washington,
Department of Plant Biology,
Stanford, CA 94305, USA
Fax: (650) 325-6857
E-mail: rhee@acoma.stanford.edu

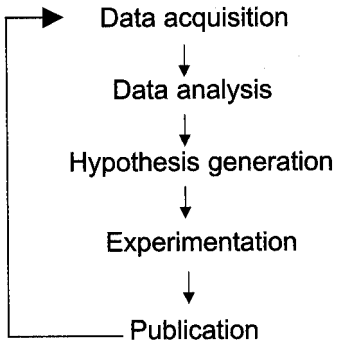

Figure 1. The *modus operandi* of biological resaerch i.e. the scientific method

With the release of the genome sequence of two flowering plants, a fungus, fruit fly, worm, mouse and the draft sequence of human, biology has entered a new era (AGI, 2000; Adams *et al.,* 2000; Venter *et al.,* 2001). What exactly is this new era and what does it signify? The *modus operandi* for biological research, as shown in Figure 1, remains the same but each step of this procedure has been undergoing significant changes in the last ten years. The use of the Internet, databases, software, and engineering to help carry out, analyze, and publish biological information has been changing the way we conduct, and think about, research.

In the traditional approach to research, the researcher first gathers information from the literature and from direct communications with colleagues. The information is synthesized to generate hypotheses and define experimental methods to test them. If there is enough information to support (or in some rare occasions reject) the hypothesis, the data are assembled and published in a scientific journal. The research community uses this published information to acquire information for their own hypothesis generation and experimental design. This cycle can continue for 20-30 years in a researcher's professional lifetime, and through this refinement, he or being an 'expert' in the subject.

One of the major changes that has occurred in the past ten years is the production of data on a large scale. The sequencing of whole genomes is a prime example. More recently this type of data production has been expanded to include genome-scale expression studies and two-hybrid information, and is expected to evolve into the production of more complex data types in the future (Figure 2). Another major change is the format of publication. These large data sets are published in a relatively raw form, without the context of a hypothesis, and often without the integration of the results into the larger framework of existing knowledge. These data are generally published in databases rather than journals, and are made accessible to researchers through the Internet. Researchers can access the data but in relatively rudimentary and simple ways. Both the increased quantity and availability of data in relatively raw format provides a challenge to scientists who must analyze the data for hypothesis generation and experimental design.

In order to derive meaning from large sets of data and build hypotheses to test the rules of biology, we need tools to analyze the data and identify patterns within the data. Some examples of such tools (some already developed and in wide use) are listed in Figure 2. In order for the tools to be developed and refined, we need easy access to data that are structured in a way that it is readily amenable to analysis. Furthermore, the analyzed data must be made accessible in order to refine and verify the analysis, identify areas poorly understood, and build hypotheses to test the correlation made from analyzing the data. The area of research that connects all these stages is information management. It is an integral part of the scientific method in this era and its successful implementation will be critical during the transition into the new era of biology. she ends up

There are many examples of databases and projects that are trying to address the challenges of the biological

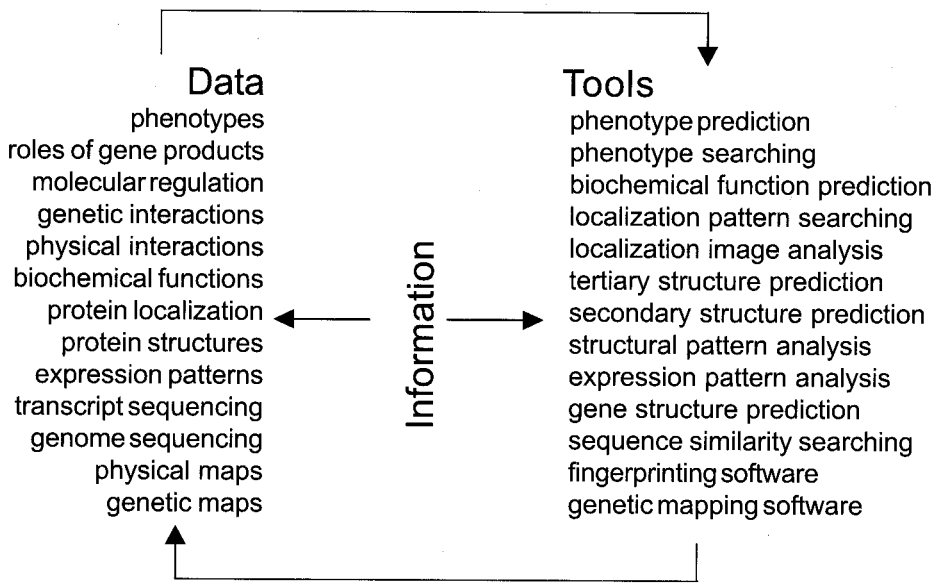

Figure 2. The relationship between large-scale data acquisition, bioinformatics tools development and information management.

information glut today. These projects can largely be divided into four major types: lab management tools, databanks of single data types, data warehouses of multiple data types, and data mining tools databanks and data workshop2010.html). warehouses. Some examples of each of these types of resources are listed in Table 1.

Nucleic Acids Research publishes a special issue on databases in the beginning of each year, which provides nice synopses of the available information management systems (NAR, 2001). Here, I will focus on describing one example of a data warehouse project, the *Arabidopsis* Information Resource (TAIR), for a model plant species, *Arabidopsis thaliana*.

Arabidopsis thaliana is an annual plant, belonging to the mustard family. Its small genome and ease for genetic and molecular manipulation, among other researchers to study this plant (TAIR, www.arabidoosis.org/info/ aboutarabidoosis.html). Its genome completed at the end of last data types for *Arabidopsis* (Huala *et al.,* year (AGI, 2000) and projects are underway to study the global expression patterns and creation of null mutants for every gene in the genome (AFGC, www.afgc.stanford.edu). In the next ten years, researchers will systematically work to identify the function of each gene product, driven by an NSF initiative termed the 2010 project (www.arabidopsis.org/info/ to search the different For researchers to be able to address the functional characterization of every gene in this organism, it is essential that all the information known for this plant be accessible in an easy, unambiguous, and intuitive way. It is also essential that the data produced by these initiatives be fed back into the system to drive the research forward. One of the key factors for accessing data from such a database lies in the importance of data association. This allows the identification of new correlations. However, for the researchers to formulate the hypotheses from the correlations in a meaningful way, it is critical that the data be traceable, both to the contributor of the data as well as to the method used to produce/ analyze the data. qualities, have motivated many The Arabidopsis Information Resource (TAIR) is one of the examples of information management systems that provides different sequence was 2001). The basic structure of the database includes relationships among data objects (clones, genes, sequences, genetic markers, polymorphisms, transcripts, etc.), description

Table 1. A few examples of Information management resources. TAIR: the Arabidopsis Information Resource, SGD: Saccharomyces Genome Database, SRS: Sequence Retrieval System, ISYS: Integrated System.

Types	Examples	URLs
Lab management Tools	BioNavigator	www.bionavigator.com
	DoubleTwist	www.doubletwist.com
	Biowire	www.biowire.com/bw_jsp/home_top.jsp
Single data type Databanks	GenBank	www.ncbi.nlm.nih.gov/Genbank/index.html
	Protein Data Bank	www.rcsb.org/pdb/
Multiple data Warehouses	TAIR	www.arabidopsis.org
	FlyBase	flybase.bio.indiana.edu
	SGD	genome-www.stanford.edu/Saccharomyces
Data mining Tools	SRS	srs.ebi.ac.uk/
	ISYS	www.ncgr.org/research/isys/

or annotation (function, map position, expression, etc.), and attribution (source of data, update history, and references) of the data objects. In this relationship, the data objects are associated to their descriptions and associated to other data types by their common and/or different descriptions. The basic searching,
browsing, updating, and downloading the data. done from a
variety of methods and perspectives, including

a graphic visualization and alignment tool for analyzing different map information (TAIR MapViewer: www.arabidopsis.org/servlets/ mapper) and textual searching interfaces for

different data types (TAIR DB Search: www.arabidopsis.ora/search). Another important aspect of data association lies in the tools to describe the data in common ways using controlled vocabulary. Toward these

ends, different database groups are developing

a set of controlled vocabulary to name and www.aeneontologv.ora).

It is tremendously exciting to think that we are at a stage where we can start asking questions about what is known, and more importantly,

what is unknown, in more systematic ways. We are becoming bolder as a research community in asking those questions. Researchers in the plant community are coming together in

discussing these issues (www.arabidopsis.org/

Systematic approaches to research will take us to that goal in large steps, but making sense of the results of systematic research programs with the hypothesis-driven, reductionist approach will be required to create knowledge from the information. functionality of data access includes

Acknowledgements Searching and browsing can be

I am grateful to Leonore Reiser, Lukas Mueller, Eva

Huala, Marga Garcia-Hernandez, Dan Weems, Chris Somerville, Tim Littlejohn, Barbara Jasny, and Pam Hines for stimulating discussions and helpful comments on the manuscript. SYR is supported by the National Science Foundation grant # DBI-

9978564. This is Carnegie Institution, Department of Plant Biology publication number 1487.

References:

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis*

thaliana. Nature. 408: 796-815. describe data objects (GO: Adams *et al.,* (2000) The genome sequence of *Drosophila melanogaster.* Science. 287: 2185-95.

Venter et *al.,* (2001) The Sequence of the Human Genome. Science. Nucleic Acids Research. 291: 1304-1351.

Huala *et al.,* (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids

Research. 29:102-105. info/carnegieworkshop.html).