Published in final edited form as:

Bioscience. 2009 February 11; 59(2): 113–125. doi:10.1525/bio.2009.59.2.5.

Biological Resource Centers and Systems Biology

Yufeng Wang [Assistant Professor]¹ and Timothy G. Lilburn [Senior Research Scientist]²

Yufeng Wang: yufeng.wang@utsa.edu; Timothy G. Lilburn: tlilburn@atcc.org

¹Department of Biology, University of Texas at San Antonio, 6900 North Loop, 1604 West, San Antonio TX 78249, United States

²Department of Bacteriology, American Type Culture Collection, 10801 University Boulevard, Manassas, VA 20110, United States

Abstract

There are hundreds of Biological Resource Centers (BRCs) around the world, holding many little-studied microorganism. The proportion of bacterial strains that is well represented in the sequence and literature databases may be as low as 1%. This body of unexplored diversity represents an untapped source of useful strains and derived products. However, a modicum of phenotypic data is available for almost all the bacterial strains held by BRCs around the world. It is at the phenotypic level that our knowledge of the well-studied strains of bacteria and the many yet-to-be studied strains intersects. This suggests we might leverage the phenotypic data from the data-poor bacteria with the omics data from the data-rich bacteria, using our knowledge of their evolutionary relationships, to map the metabolic networks of the little-known bacteria. This systems biology-based approach is a new way to explore the diversity harbored in BRCs.

Keywords

systems biology; biological resource centers; networks; metabolic maps; diversity

Biological resource centers (BRCs) are institutions that store and maintain the subject materials of biological research, and provide services related to these materials. They also collect and store data and information relevant to their holdings. BRCs vary in size and can specialize in certain areas of biology, provide support and services to their customers, as well as conduct research on their holdings, frequently in collaboration with scientists at other institutions. BRCs have existed since the 16th century, when Luca Ghini established a botanical garden and dried specimen collection, or herbarium, at the University of Pisa in Italy (Pavord 2005). The purpose of the collection was to make specimens available for study, especially specimens of plants that were rare or difficult to obtain. Ghini shared his dried specimens with other scientists, thus establishing the use of specimens as a means of sharing information. This innovation was the first step down the road to the discipline of taxonomy and enabled the foundation of a universal nomenclature for plants in Europe.

In the centuries that followed, other herbaria and collections of all kinds of biological specimens were established. Some BRCs are primarily archival repositories and distribution centers for organisms. A few large BRCs offer a diversity of organisms as well as services and information relevant to the organisms. Ghini's herbarium contained some hundreds of specimens; today, the herbarium of the Muséum National d'Histoire Naturelle in Paris contains roughly seven and a half million specimens. Due to their small size, microorganisms were among the last

groups of organisms to be gathered in public service collections. They had been observed for the first time in the 17th century and the first species were named in the 18th century (Müller 1773). By the late 19th century, František Král at the German University of Prague (Uruburu 2003) had established the first collection of microorganisms. Other collections soon followed. In 1891 the Collection de l'Institut Pasteur (CIP) was founded, in 1894 the Mycothèque de l'Université Catholique de Louvaine and others followed. The American Type Culture Collection (ATCC) was founded in 1925 with about 175 strains of bacteria (Hay 2003). Interestingly, about half the microbes from Král's first collection eventually made their way to the ATCC, where they can still be found today along with 18,500 strains of Bacteria and Archaea and 45,000 fungal deposits. Today there are 531 microbial collections registered with the World Federation of Culture Collections; they contain over 560,000 bacterial deposits and over 460,000 fungal deposits (figure 1).

Many of the microorganisms held by BRCs are important to research activities in fields ranging from environmental microbiology to medicine. They also find direct application in industrial settings in the production of food, pharmaceuticals, and fragrances and in enhancing agricultural processes, either as whole organisms or as derived products, such as enzymes. Living organisms are used to produce many products. The most famous is probably yeast, *Saccharomyces cerevisiae*, which is used to produce wine, beer and bread. Other microorganisms are used to produce vinegar, soy sauce, antibiotics, vitamins and a host of other commodities. Enzymes derived from microorganisms enhance processes in the food and chemical industry. Enzymes are incorporated in detergents, used to treat textiles, employed in the production of beverages and dairy products and are increasingly used to replace less environmentally friendly processes once used to modify seed oils, tan leather and process natural rubber.

The astounding abilities of fungi and bacteria to breakdown or transform both natural and man-made compounds is being exploited to clean up contaminated soils and water. Bioenergy researchers also seek to use these abilities in the production of ethanol, alkanes and hydrogen.

Although the usefulness of bacteria and their derivatives is widely recognized, the number of strains that we actually take advantage of is relatively small compared to the number held in BRCs. The holdings of BRCs contain unrecognized and therefore untapped resources for medical, agricultural and biotechnological applications. This potential has remained unrealized because, at bottom, we have a very limited idea of what most of these strains can do. Typically, a strain is deposited either because it is a type strain, that is, the strain of a species of bacteria that establishes the characteristics for the species, or because it had some other property that made it interesting to researchers at the time. Over the years and decades following the deposit of strains in a collection, data and information pertaining to these strains accumulates in the published literature, in public databases and at the BRC, but the information appears to be distributed according to a power law distribution - we have a lot of information about a few strains, but not much information about most. This is illustrated in figure 2, which shows, in panel (a), the numbers of NCBI gene identifiers (GIs) linked to each of the properly formed (Latin binomial) names in the NCBI taxonomy for the Bacteria and Archaea and, in panel (b), the PubMed Central articles linked to the same set of names. While this is only an approximation of the knowledge we have about the organism represented by each name (for example, genome sequences are frequently represented by only one or two GI numbers), we feel it accurately reflects the real situation. Most research efforts center on a few organisms. Although there are currently 499 complete genome sequences associated with fully named organisms, most of these organisms are not intensively studied and much of the annotation supplied with the sequence data has been transferred from other, better-studied organisms on the basis of sequence similarity. For the great majority of deposited cultures the accessible knowledge is represented by a single sequence and/or one or two relevant publications. One

publication usually describes the bacterium and outlines the characteristics of interest. For some deposits there is no readily available information. For example, *Streptomyces albosporeus*, a member of a genus well known as a source of antibiotics and for its ability to carry out chemical transformations, has no PubMed hits, and cannot be directly linked to any nucleotide or protein sequences, 91 years after its first description and 79 years after its deposit at ATCC. Information was published in 1916, but is not cataloged on PubMed. Researchers who are aware that *Streptomyces albosporeus* is a synonym of *Streptomyces aurantiacus* will find three PubMed citations; two of these deal with the classification of *Streptomyces* species and one deals with properties of this organism. BRCs that have this organism do hold some relevant data, the results of phenotypic tests used to characterize the strain and ensure the stability of the organism during decades of storage, but these data are not generally available outside the BRCs. In summary then, less than 1% of named Bacteria and Archaea have a significant number of publications associated with them. The amount of available sequence data is larger – about 8% of named Bacteria and Archaea have had their genomes sequenced, but this still leaves a lot of uncovered territory.

Here, we propose a route to uncovering the enormous metabolic power represented in the collections of microorganisms (specifically the Bacteria) held by BRCs, using an inferential approach. Our proposed route adopts the systems biology methodology. The recent recrudescence of interest in systems biology has led to the development of the tools necessary for exploring and then exploiting the hidden resources in BRCs.

Systems Biology

Systems biology posits that an organism is more than the sum of its parts, that is, that the emergent properties that constitute an organism's physical characteristics cannot be explained by an examination of all the individual components that constitute the organism. The study of organisms in a holistic way is not new; as Kahlem and Birney point out (Kahlem and Birney 2006), it was the only way to study organisms until relatively recently. In the 19th century, scientists began to take organisms apart at the molecular level to study their component parts. As the central dogma of biology was elucidated, increasingly sophisticated methods for studying molecular biology were developed and in the latter half of the 20th century molecular biology dominated many fields of biology. With the advent of the genomics revolution it has become possible to assemble a complete list of the genes and gene products for an organism -- in essence a parts list. The subsequent development of other omics approaches has allowed us to gather further data from experiments designed to detect whether and under what conditions the components appear in an organism. The result has been an almost overwhelming number of data. Within the last 10 years, it has begun to become more widely recognized that the systems biology paradigm offers a way to extract meaning from these data, and to go beyond merely taming the data to using them to construct comprehensive views of an organism, models that can have enormous explanatory power.

Systems biology, as a way of thinking about and exploring biology, traces its roots to the work of Ludwig von Bertalanffy on systems theory (Bertalanffy 1969) and Norbert Wiener's work on cybernetics (Wiener 1948). Subsequent to the publication of that work, biologists began to apply "systems thinking" to biology and sought to uncover the basic laws governing evolution and behavior, or to at least formalize what were (and are) intuitive explanations of biological phenomena (for a review, see Wolkenhauer 2001). The construction of models is central to systems biology. Initial successes in systems biology modeling centered on the biochemistry of metabolism, with modeling techniques developed as part of Metabolic Control Theory (Heinrich and Schuster 1996) and Biological Systems Theory (Voit 2000). However, the full development of systems biology was impeded by a lack of data. Therefore, the genomes and allied masses of data generated by high-throughput methods that began to appear in the 1990s

represented an opportunity. If integrated carefully, these data can be used to build a model that represents a systems-level understanding of a given piece of cellular machinery, or a cell, or an organ or an organism and so on, depending on the type, quantity, and quality of the data and the level of abstraction desired. In addition, the more data that can be integrated into the model, the more reliable is our view of the system, as any noise in the different data sets will tend to cancel out and intensify the signal. Another important factor in the emergence of systems biology has been the constant increase in the power of desktop computers. For an overview of systems biology see (Aderem 2005, Barabasi and Oltvai 2004, Cornish-Bowden 2006, Hood et al. 2004, Ideker et al. 2001, Kitano 2002, Xia et al. 2004).

Using systems biology to explore BRC holdings

As mentioned above, the bacteriology collection at ATCC, where one of us is employed, is typical of large BRCs in that it contains many thousands of strains of Bacteria and Archaea, of which only a few hundred species, representing perhaps a few thousand of these strains, are widely studied. We do not have a comprehensive understanding of the metabolic capabilities of any of them. Even the laboratory workhorse, *Escherichia coli* K12, continues to surprise investigators with new abilities (Loh et al. 2006). For most strains we have a published description, at least one set of characterization data, and a name through which we can discover that strain's closest relatives. We will refer to these strains as the "data-poor" strains. For a relatively small number of other strains, we have much larger volumes of data. If we take a genome sequence to be the minimum requirement for inclusion in a set of "data-rich" strains, at the time of writing we have 787 data-rich strains that represent 499 different species. Admittedly, for some of these strains the genomic sequence and associated annotation are almost all we have, but other data, from gene expression experiments and proteomic investigations to published literature, do exist for many of these strains. All of these data are relevant to the nearest relatives of the sequenced strains; the relevance declines as the evolutionary distance between a given pair of species increases. We propose to leverage the data we do have for data-poor strains with the data from the data-rich strains using our knowledge of the evolutionary relationships among the Bacteria. This process has become routine in comparative genomics and is known as phylogenetic transfer.

There are two major phases in using systems biology to explore a BRC collection: firstly the construction of robust metabolic network models for data rich organisms and secondly the extrapolation of these models to data-poor Bacteria for which we have only phenotypic data. In the following sections, we will discuss the data, tools and issues, including potential obstacles, relevant to the proposed approach, including a discussion of how networks for data-poor Bacteria will be inferred.

Network model construction can be conceptualized as four iterative steps (1) the harvesting and processing of data, (2) the integration of data, (3) the development of a network model and (4) the validation of the models (figure 3).

Harvesting and processing data from well-studied, well-understood organisms

The available data can be classified as either genotypic or phenotypic. Genotypic data are derived from gene sequence analysis and other types of analyses that examine gene content, gene expression and so on. Almost all the data springing from the genomics revolution are, or result from, the analysis of, genotypic data. Phenotypic data consist of qualitative and/or quantitative characteristics of an organism. These characteristics include morphological traits (e.g., size and shape), physiological traits (e.g., salt tolerance), and biochemical traits (how they obtain the energy and carbon needed for growth).

Genotypic data

Modern high-throughput approaches to biology are generating unprecedented volumes of data. These data yield information about different levels of biological organization. Genome sequences, perceived as the fundamental level of organization, can yield a list of genes, and analysis of the genome can reveal which genes are in the same location, which appear to be co-regulated, which are responding to selective pressures and which are under purifying selection. As the quality and quantity of annotation varies from one sequencing project to the next, it is important to try and normalize the data as much as possible. Secondary sources for genomic data, such as the JCVI Comprehensive Microbial Resource or the EBI's Integr8 database re-annotate the genomes in a uniform way, so using genomic data from these sources improves the comparability of the results. Other databases, such as the USDOE's IMG database can also help ensure the quality of the genomic annotation data.

Genome sequences also yield lists of encoded proteins and this naturally leads to data that looks at gene expression. Microarray and SAGE (serial analysis of gene expression) experiments reveal which genes are actually transcribed and mass spectroscopy experiments or two dimensional gel studies can tell us which proteins are actually translated. These data can also provide evidence as to which elements are co-regulated and are often correlated with changes in phenotype. Functional information about the encoded proteins that goes beyond what is included in the genome annotation, is available from a multitude of databases. For metabolic modeling the enzyme information databases, like BRENDA, are very useful. More generally, two of the most popular databases are UniProt and InterPro. Table 1 shows information on the resources and software mentioned in this review. A wider view of these types of resources is available from the annual issues of *Nucleic Acids Research* that are devoted to database and web resources.

The third level of data reveals interactions among these components. Pathway databases (such as MetaCyc, KEGG, and Reactome) present the set of interactions that drive the metabolism. These databases have different concepts of what constitutes a pathway: MetaCyc-based pathway genome databases (PGDBs) are specific to an organism, while KEGG pathways are more general. Protein-protein interactions can be retrieved from databases like the STRING database, which estimates interactions among proteins using information derived from their genomic context (are they in the same regulatory unit? Are they frequently found on the same genomes?), from seven other protein interaction databases, from the literature, and from the results of transcriptomics or proteomics experiments and other high-throughput omics experiments. The probability that two proteins interact is estimated and a confidence value is attached to all the interactions found for a given organism. The map of interactions (figure 4) is quite different than metabolic pathway maps (figure 5) because all kinds of interactions, physical and functional, are shown. It is difficult to trace a biochemical pathway within the interaction network, but the network reveals interactions not touched on in the metabolic maps and is a good way to see an overview of all the annotated proteins in an organism. Finally, this level of data includes data from interactional analyses, which are drawn from observational studies that collect data about the organism under some condition or from perturbation experiments that induce changes in the system, for example by deleting a gene or otherwise suppressing its expression, and characterize the change in gene expression. All of these data can be generated in a high-throughput way and all can enhance a systems-level view of an organism.

The phenotypic data

are derived from our observations of the behavior of organisms. They are relatively sparse, but, as we shall see, they assume great importance in our approach to exploring BRCs because the information we have about the data-rich and the data-poor strains intersect at the phenotypic

level. The data we are most interested in here are the results of a suite of tests that are widely used to identify Bacteria and, in BRCs, to authenticate the organisms when new seed stocks are grown up. Given the importance of authenticating both newly accessioned prokaryotes and distribution cultures grown from seed stocks, these routine phenotypic characterization tests used have been rigorously standardized.

The three types of phenotypic tests most commonly carried out are the inhibitory, biochemical, and nutritional tests. Inhibitory tests measure a strain's resistance or susceptibility to compounds such as antibiotics, dyes and heavy metals. Biochemical tests assay an organism for specific enzymes and pathways. Nutritional tests examine an organism's ability to grow on a substrate as a sole carbon, energy and nitrogen source, for example, or the organism's requirements for vitamins or other growth factors. Underlying each biochemical, nutritional or inhibitory test is one or more metabolic capabilities of the organism, but the nutritional tests are perhaps the most informative, since they test for growth.

Currently, such phenotypic data are obtained from two streams. The first stream is from the work done to identify and maintain bacterial strains in clinical labs and in BRCs. Typically, these data are generated from the classic characterization schemes for the identification of bacterial strains. These schemes constitute a suite of phenotypic tests, most commonly carried out on two platforms. In the first, widely used approach the tests are carried out in test tubes and on agar plates. The evaluation of the tests relies on the judgment of a biologist. As some tests are not reliably reproducible and judgments about tests can sometimes be subjective, there is usually a 5% error rate in scoring these tests. One very favorable aspect of the nutritional tests when done this way is that the entire metabolic network is tested – the organism must be able to grow on the tested compound. Packaged single use identification kits constitute a second type of test platform. The first of these to appear was the API 20E strip for the identification of *Enterobacteriaceae* (bioMérieux Inc., Marcy l'Etoile, France), which reported the results of 21 biochemical and nutritional tests. Variations on this kit now target 550 species. Other systems are also available, some machine-readable, but they all share two weaknesses. Firstly, these packaged tests are intended for use in clinical microbiology labs and they may not be relevant to strains of bacteria that are not of clinical interest, either because the included tests do not detect traits of these strains or because the reference database used to make an identification does not have data for these strains. Secondly, it is not always clear what these proprietary tests are measuring.

BRCs have phenotypic data (sometimes called characterization data) for most of the bacteria they hold. At the ATCC, phenotypic data have been gathered for over 80 years. These data are not made publicly available by any BRC that we are aware of, probably because the costs of mobilizing the data are very high. The literature and specialized publications like Bergey's Manual of Systematic Bacteriology (Garrrity 2001), contain a vast amount of this type of data, but it is not suitable for high-throughput use. GIDEON Informatics Inc. (Los Angeles, CA) offers a commercial source for phenotypic data that covers 900 bacterial taxa, all disease-causing bacteria or human commensals.

A second source for phenotypic data is from the research community. In the 1960s the advent of computer-based schemes for the identification of bacteria, which were based on the principles of numeric taxonomy (Sneath and Sokal 1973), heightened researcher's interest in the development of characterization schemes. It is possible to find some of these data online (for example, <https://www.soton.ac.uk/staff/tnb/pibnew2-.htm>), but the coverage of bacterial diversity is patchy. Research in this area led to the development of instruments like the API 20E system and by the 1990s published research on this topic had fallen off considerably as the field became commercialized. Recently, research interest in phenotyping has rekindled and high-throughput methods for characterizing the phenotype have been

developed. One such system is the Phenoplate (Bactus AB, Huddinge, Sweden), which allows up to 48 different biochemical tests to be rapidly evaluated. The Biolog OmniLog Phenotype Microarray system (Biolog Inc., Hayward, CA) (Bochner et al. 2001) has much greater potential as a high throughput phenotyping system. The OmniLog PM comprises a set of 96 well plates in which each well tests for a different aspect of cellular metabolism. Up to 1,900 tests can be carried out in a high throughput way, using a dedicated incubator and plate reading system that measures the kinetics and extent of substrate utilization. The OmniLog PM system is a relatively novel technology and researchers are still learning its uses, but more publications demonstrating the value of the data produced by this machine are appearing (see, for example, Loh et al. 2006, Mols et al. 2007, Oh et al. 2007). One disadvantage of this technology compared to the classic approach is that it does not test the complete metabolic network. Instead, it detects changes in redox potential in the growth medium. Thus, a bacterium that can partially metabolize a substrate could be scored positive in this system, even though it is not capable of growth on the tested substrate. Also, while the developers of this instrument have tried to ensure that the media and tests used are relevant to as broad a diversity as possible, users should be aware that the system may not work as well for some taxa as for others.

The results of the phenotypic tests are evidence for the existence of among other things, specific enzymatic activities or metabolic pathways and, therefore, for the existence and expression of a gene or genes in the tested organism. A positive result for a test in a data-poor organism gives us some insight into the genetic potential of the bacterium and this potential can frequently be used in other ways to produce other phenotypes. This is the critical assumption in the use of systems biology methods to exploring the metabolic potential of a bacterial BRC. As an example, let us look at one phenotypic test, the N-acetyl-D-glucosamine as sole carbon source test. For a bacterium to be able to use this substrate it must be able to convert the N-acetyl-D-glucosamine to fructose-6-phosphate, an intermediate in the central metabolic glycolysis pathway. Figure 4, Figure 5, and Figure 6 show visualizations of the interactions among the components involved in the N-acetyl-D-glucosamine degradation pathway of *Vibrio cholerae*, along with the downstream components required for growth on this compound. The proteins and other components involved in the N-acetyl-D-glucosamine degradation and glycolysis pathways have been highlighted in each figure. Figure 4 and Figure 6 vividly underscore the fact that the components of these pathways are parts of many other interactions and processes in this organism. For example, the proteins involved in the N-acetyl-D-glucosamine degradation pathway are linked, either by co-transcription and expression or by shared substrates, with proteins that are part of five other degradative or biosynthetic pathways. Thus, the test for growth on this substrate implies the presence of not only the direct, downstream central metabolic pathways, but also other pathways based on the likelihood that the enzymes and compounds tested for are available for other purposes.

After falling out of favor during the molecular biology era, phenotypic data is re-emerging in the post-genomic era. This renewed interest is driven, at least partially by the reemergence of systems biology and its goal of linking genotype to phenotype. The combination of genomic and phenotypic data has yielded insights that were not available from the genomic data alone, probably because a large proportion of the genes in a new bacterial genome are of unknown function. For example, Perkins and Nicholson (Perkins and Nicholson 2008) looked at the effects of *rpoB* mutations on the metabolic capabilities of *Bacillus subtilis* and found that certain mutations unveiled hitherto unrecognized capabilities in this bacterium, specifically the ability to metabolize D-psicose, gentiobiose and β -methyl-D-glucoside. The enzymes that constitute the pathways involved in metabolizing these compounds were not recognized in the genome annotation process, but the capability was revealed when the *rpoB* mutations interfered with the normal regulatory suppression of these pathways. In a similar way, Jones et al. (Jones et al. 2007) used phenotypic profiling to explore the metabolic processes controlled by the RpoN sigma factor in *Pseudomonas fluorescens* SBW25. Another use of phenotypic data is to

examine the links between genotype and phenotype by looking for statistical correlations among phenotypes and proteins, as represented by the proteins pfam functional classifications. In the Gerstein lab, correlations between phenotypes and functional groups of proteins, as grouped according to their pfam classification, or their Gene Ontology classification, or their presence in a KEGG pathway (Liu et al. 2006) were searched out, and the correlations (and anti-correlations) were shown to have explanatory power, supporting a genotype-phenotype linkage. This group then went on to show that the results from the types of phenotypic tests typically done by BRCs can be used to predict the presence or absence of a protein in a bacterium (Goh et al. 2006). Research linking phenotype to genotype via literature mining has also demonstrated the ability to predict the presence of unrecognized proteins in bacteria (Korbel et al. 2005). Although more and more papers that make use of phenotypic data are appearing, there are no instances of integration of phenotypic data with other data to predict metabolic networks, using the type of unified modeling framework discussed below.

Data integration

Data integration involves combining heterogeneous data from different sources in order to better understand the object or system to which the data pertain. The promise of integration is that it will allow us to extract more information from the combined datasets than we could if the datasets were considered individually. This is because the “noise” in the data sets should cancel out, being random, while the “signal” should be reinforced. Searls has recently published an excellent review of data integration (Searls 2005). There are potential obstacles to data integration that must be considered before data integration can begin. Firstly, the data must be consistent and comparable. Gene expression levels, for example, may have been measured under different conditions or on different platforms. The identifiers or names used for data objects could differ between data sets. Such problems generate conflicting results. Secondly, the accuracy of the data must be known or measurable. Uncertainties about the data sets include the extent of variation within the datasets and the causes of the uncertainty in each data set. We know there are in-built uncertainties in data associated with measurement. Disparities of scale (consider the span of life’s evolution versus the rate of electron transfer), sample variation and non-linearity introduce variation that cannot be measured and there are also uncertainties that result from our lack of knowledge about the systems we are studying, for example, in the underlying causes (Joyce and Palsson 2006, Reed et al. 2006). Obviously, it is important to take into account the characteristics of the data, where they came from, and what their intended use was when we plan which data we will be using to build our systems level views and what tools we will use to build them.

We can conceptualize three methods for data integration. The first and simplest method for data integration is the creation of a database. Databases create links among data and when properly modeled, allow queries that reveal meaningful relationships to be made.

The second widely used way to integrate data is founded on databases as well, but provides a framework for the database(s) to aid users in visualizing and interpreting the data. This approach is adopted by the genome sequence databases, whether they are general collections of genomes or organism-specific databases like PlasmoDB. The BioCyc pathway genome database collection is a good example of this type of data integration; they center on representations of the metabolic networks of included taxa.

The previous two methods of data integration create links among the data based on the database model used. The links between two data items may be erroneous (linked due to experimental error), coincidental (co expression for fundamentally different reasons), of low reliability (the data items are mentioned in the same literature abstract), or incontestable (the system that has been intensively studied in the wet lab). Statistical approaches to data integration have been adopted because they can accommodate these kinds of variability in the data, uncertainty,

missing data, and so on. The statistical approach to data integration is exemplified by Pointillist software package, developed at the Institute for Systems Biology (Hwang D. et al. 2005a). With this software, users can apply different weighting methods and decision boundary determinations to determine the most appropriate way to integrate disparate data types. The results are displayed in a network model that captures the interactions among the data elements and includes confidence estimates for the existence of a given node and the interaction between a pair of nodes (Hwang Daehee et al. 2005b). Troyanskaya et al. (Myers and Troyanskaya 2007, Troyanskaya et al. 2003) and Tanay et al. (Tanay Amos et al. 2004, Tanay A. et al. 2005) have developed other methods, which also present the data integration results in the form of a network. In fact, this concomitant data integration and network inference points to the use of probabilistic graphical modeling frameworks for data integration and construction of network models and we will discuss this further below

Construction of network models

Network models are usually visualized using a graphical notation that describes the relationships among the elements of complex systems. In its simplest form, this notation uses nodes (or vertices) to represent genes or proteins and edges (directed or undirected lines) to represent the interactions between a pair of nodes. Models can also be stored in a machine-readable format, such as the Systems Biology Markup Language (SBML) or BioPAX, that enables exchange of models among different software tools. There are two fundamental ways to construct network models. The first uses mechanistic modeling and has emerged from the field of biochemistry and modeling of biochemical pathways. The actions of enzymes are well studied and it is possible to model their activities using a series of equations. The most detailed mechanistic approach is kinetic modeling. Kinetic modeling takes advantage of detailed information, often gathered over a period of years, about the system under study. The results are low-level models that are highly detailed, but the complexity of the modeling approach means that the modeling process becomes intractable as the system being studied increases in size. Voit (Voit 2000) cites the example of glutamine synthetase; eight reactants and modifiers affect this enzyme. If we wanted to establish the rate law for this enzyme we would need to carry out about 100 million assays, and the rate law would consist of about 500 terms (Savageau 1976, Woolfolk and Stadtman 1967).

Constraint-based methods (Price et al. 2004), another mechanistic approach, deal with some of the complexity of mechanistic methods by imposing stoichiometric, thermodynamic, and enzyme capacity constraints on the range of possible solutions. The method starts with a model of the network being studied, which, again, is usually the result of many years of study of biochemical pathways. The method does not find a single solution, but a range of possible solutions that are consistent with both the constraints and the observed behavior of the network (the phenotypes). Finally, the solution that most closely approximates the actual network behavior is found, usually in the context of optimizing some factor such as growth rate (Price et al. 2003, Reed and Palsson 2003). Constraint-based approaches have been very successful in modeling metabolic networks and the resulting models have been able to predict accurately the results of gene deletions in *E. coli*, for example (Feist et al. 2007, Fong and Palsson 2004). Recently, efforts to incorporate regulatory information into the models have also met with success (Covert et al. 2004).

The second way of building network models uses probabilistic approaches. They have been adopted because they take advantage of statistical treatments that allow us to deal with uncertainties and can be combined with methods for integrating the disparate data types from which the network is inferred. Probabilistic approaches can yield what Ideker and Lauffenburger (Ideker and Lauffenburger 2003) classified as “high-level abstractions” of the actual networks in an organism. These abstractions give information about the components

and connections within the network and take the form of a network visualization (Friedman 2004). The network model is, of course, a summary of the data used to build it. But it is also a collection of predictions or hypotheses about the system under study and may contain novel, unforeseen information about the system and about the relationships between the entities we are modeling. If methods that can handle more information are used, such as Bayesian or Boolean models, then network influences and information flow within the network can be traced. Unlike mechanistic approaches, probabilistic methods can generate networks *de novo*; in this they are analogous to exploratory data analysis, since the networks are based entirely on information that is in the input data.

The development of probabilistic modeling methods for biological network inference is a very active field of research (Huang et al. 2008). Models investigated range from a simple, fully correlated model (Beal et al. 2005, Friedman et al. 2000, Pe'er et al. 2001) that is very tractable, but does not deal well with noise, to a variety of Gaussian models that attempt to deal with noise and randomness. Among the Gaussian models we find graphical Gaussian models, which use partial correlations that allow us to distinguish, for example between direct and indirect interactions between two genes (Schafer and Strimmer 2004, Toh and Horimoto 2002). Linear Gaussian modeling has been popular, but it fails to adequately deal with the non-linearities in the dependencies within a network. Gaussian models that are adjusted to accommodate nonlinearity have been proposed. Probabilistic Boolean models have also been explored (Shmulevich et al. 2002a, Shmulevich et al. 2002b). These deal with the essentially deterministic nature of Boolean models by focusing on a set of Boolean functions (or predictors) for each data object and combining them into a probabilistic model. This adds randomness, and the model resembles a discrete Bayesian network model (Lahdesmaki et al. 2006). Having chosen a model suitable to the data at hand, the next step is to infer networks from the data. One of the most fruitful approaches to network inference takes advantage of probabilistic modeling techniques. This approach combines graph theory and probabilistic theory. It allows the complex global models to be built up from simple local models and ensures that the model is extensible, allowing it to account for additional aspects of the system or new datasets. Much of the work on probabilistic graphical modeling has been developed in the field of signal processing (Kschischang 2003) to deal with problems seen in wireless communications, antenna array processing, image processing, and biomedicine. Frequently the problems can be generalized as signal versus noise problems and, as we have seen, this is analogous to the problem of extracting signal from noise in more or less disparate data sets. In signal processing, this problem has most efficiently been dealt with using Bayesian approaches. Bayesian signal processing is a special case of the graphical modeling approach. It has a long and successful history in modeling complex systems and we propose that its methodology that can be used to better understand biological networks as well. Recall from the section on data integration that the usual means of displaying the results of statistical data integration is as a network. In fact, the natural product of the data integration is a network of relationships that reflects how the elements in the system interact and presents insights into how the system functions and it is here that Bayesian methods of probabilistic graphical modeling shine, because Bayesian methods allow prior information to be injected into inference through the prior distributions. Thus, the inference results from older data sets (that is, the previously inferred network topology) can be treated as a form of prior information and be integrated with newer data sets. The inference from the initial network is efficiently reused and data reprocessing can be consequently avoided. This naturally accommodates the iterative nature of biological research, which demands that data integration be a sequential process and that new data be readily incorporated into the inferred network model.

Overall, there are two goals for Bayesian network inference. The first and most interesting is learning the network topology. The second, estimating the parameters of the model, can shed light on aspects of the model system, but it can be difficult to interpret their meaning. When

we wish to determine the most likely topology that is supported by the data, so-called point (or hard) solutions are sought. Probabilistic soft approaches, which provide estimates of the *a posteriori* probabilities (APPs) of the topology, allow us to estimate the confidence of inference and, most importantly, to integrate data from disparate sources.

Inferring networks for data-poor Bacteria

In order to leverage the phenotypic data available for data-poor organisms, we are using the soft probabilistic approaches just mentioned within a Bayesian probabilistic graphical framework. We have chosen this approach because it is statistically robust, accommodates different types of data and the models can be updated easily. Within this framework, data integration and network inference can be systematically performed.

The first step in inferring networks for the data-poor bacteria is the creation of a set of inferred networks for the data-rich bacteria. These will then serve as priors. To build the networks for the data-rich bacteria, we draw on all the types of genomic data discussed above, as well as derivatives of these data based on our own or other group's analyses. One important type of derived data is the logical metabolic network. This is a network that is deduced from the genomic content of a data-rich organism, using the Pathway Tools software (Karp et al. 2002). This software takes as input an annotated genome and compares the encoded proteins with the proteins in the MetaCyc database (Caspi et al. 2008). This database consists of a set of over 1,000 metabolic pathways and their constituent enzymes. If the enzymes are found in the genome, the pathway is scored as present. The deduced metabolic network tends to have more pathways than are actually present in the bacterium, but it serves well as a foundation element for inference.

Data integration and network inference for the data-poor bacteria can be summarized in four steps. The first three steps involve inferring networks for data-rich organisms and may be repeated many times in order to build up a robust set of priors for the fourth step. It is in this last step that the networks of the data-poor organisms are inferred.

1. Evidence mapping: Map the information from logical metabolic network. The logical metabolic network data carry direct and indirect evidence on connectivity. The evidence is often qualitative and deterministic binary information (point estimates) such as 'yes' or 'no'. Even when it is probabilistic, the information is in the form of confidence intervals or p-values, which are rooted in likelihood-based inference rather than Bayesian. To deal with this problem, we can focus on the probability of the existence of each edge in the network, rather than on defining the APP for every possible topology. One version of this solution is given in (Imoto et al. 2004);
2. Prior evidence fusion: This problem is fundamentally one of aggregating information about the system. For example, microarray data might be binned according to the level of expression, or information about the correct identity may need to be included in the integration/modeling process. Many solutions have been proposed including linear pooling, logarithmic pooling, likelihood pooling, supra Bayesian pooling, and so on (Clemen and Winkler 1999, Ouchi 2004);
3. Bayesian data fusion: We can integrate characterization and other evidence data with the prior distribution from the logical metabolic network data. As stressed above, data integration and network inference must proceed hand in hand. In the graphical model arrows imply dependence between the variable sets and the direction conveys general meaning of 'conditioning' in the probability distributions.
4. Having built network models for the data-rich bacteria, we can proceed with the inference of networks for the data-poor bacteria. The process is identical to the first

three steps, except that the inputs are different and some adjustments are made to accommodate missing data. Here, the logical metabolic network is deduced from the pathways known to carry out the transformations that are revealed by the phenotypic tests. These pathways almost all intersect with the core metabolic pathways. A second significant change is that we now use a topology prior that has been adopted from a different, if closely related, species (or strain) of data-rich bacteria.

It is natural to use the core data from the most closely related bacterium when inferring a network model for the data-poor bacteria, but it is equally possible to use the core data from more distantly related bacteria to detect properties that would be considered unusual in the taxon in which the query organism is found. Of course, as the query organisms is tested against more distant taxa, the probability that a novel property actually occurs in the query organism diminishes, but it is always possible to test the query organism for the unusual property, which is the most basic form of validation.

Validation

We use two types of model validation: statistical and experimental. Statistical methods include tests against simulated networks (Jordan 1999, Kay 1993), cross validation (Roweis and Ghahramani 1999), and bootstrapping (Smyth 1997). Each method has its own advantages and disadvantages and probably no single approach can provide a reliable evaluation of the results. Drawing on our own experience, we would suggest using these approaches sequentially in order to validate the network inference algorithm and provide error analysis.

Every model makes testable predictions about the system and these can be checked in wet-lab experiments. In this case, the experiments will take the form of the phenotypic tests used to characterize Bacteria. Indeed, this type of characterization testing should form part of the iterative testing of the models as they are developed. If more than the standard phenotypic tests need to be carried out, a Phenotype MicroArray machine could be used to validate inferred capabilities. Some parts of the inferred networks may conflict with what we know to be true, but these inconsistencies with biology may actually point to the most interesting results, the discovery of novel metabolic capabilities in strains from the unexplored regions of the world's BRCs.

Acknowledgments

We thank Yufei Huang for his advice on and lively discussions of probabilistic modeling. Grant 1R21AI067543-01 from the National Institutes of Health (NIH) and the National Institute of Allergy and Infectious Diseases (NIAID) to T. G. L. and Y. W. supports this publication. Y. W. also received support from NIH/National Institute of General Medical Sciences (NIGMS) grant 1SC1GM081068-01 and NIH/NIAID grant AI080579. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the American Type Culture Collection, NIAID, NIGMS, or NIH.

References Cited

- Aderem A. Systems biology: Its practice and challenges. *Cell* 2005;121:511–513. [PubMed: 15907465]
- Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5:101–113. [PubMed: 14735121]
- Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* 2005;21:349–356. [PubMed: 15353451]
- Bertalanffy, Lv. New York: G. Braziller; 1969. General System Theory; Foundations, Development, Applications.
- Bochner BR, Gadzinski P, Panomitros E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* 2001;11:1246–1255. [PubMed: 11435407]

- Caspi R, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2008;36:D623–D631. [PubMed: 17965431]
- Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Analysis* 1999;19:187–203.
- Cornish-Bowden A. Putting the systems back into systems biology. *Perspectives in biology and medicine* 2006;49:475–489. [PubMed: 17146133]
- Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429:92–96. [PubMed: 15129285]
- Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson B. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology* 2007;3:121–121. [PubMed: 17593909]
- Fong SS, Palsson BO. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat Genet* 2004;36:1056–1058. [PubMed: 15448692]
- Friedman N. Inferring cellular networks using probabilistic graphical models. *Science* 2004;303:799–805. [PubMed: 14764868]
- Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol* 2000;7:601–620. [PubMed: 11108481]
- Garrity, GM., editor. 2nd ed.. New York: Springer-Verlag; 2001. *Bergey's Manual of Systematic Bacteriology*.
- Goh CS, Gianoulis TA, Liu Y, Li J, Paccanaro A, Lussier YA, Gerstein M. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics* 2006;7:257. [PubMed: 17038185]
- Hay, RJ. ATCC historical perspectives. In: Cypess, RH., editor. *Biological Resource Centers: Their Impact on the Scientific Community and the Global Economy*. Manassas, VA: American Type Culture Collection; 2003. p. 153-161.
- Heinrich, R.; Schuster, S. New York: Chapman & Hall; 1996. *The Regulation of Cellular Systems*.
- Hood L, Heath JR, Phelps ME, Lin BY. Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004;306:640–643. [PubMed: 15499008]
- Huang Y, Tienda-Luna IM, Wang Y. A survey of statistical models for reverse engineering gene regulatory networks. *IEEE Signal Processing Magazine* Forthcoming. 2008
- Hwang D, et al. A data integration methodology for systems biology. *Proc Natl Acad Sci U S A* 2005a; 102:17296–17301. [PubMed: 16301537]
- Hwang D, et al. A data integration methodology for systems biology: experimental verification. *Proceedings of the National Academy of Sciences of the United States of America* 2005b;102:17302–17307. [PubMed: 16301536]
- Ideker T, Lauffenburger D. Building with a scaffold: emerging strategies for high-to low-level cellular modeling. *Trends Biotechnol* 2003;21:255–262. [PubMed: 12788545]
- Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001;2:343–372. [PubMed: 11701654]
- Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J Bioinform Comput Biol* 2004;2:77–98. [PubMed: 15272434]
- Jones J, Studholme DJ, Knight CG, Preston GM. Integrated bioinformatic and phenotypic analysis of RpoN-dependent traits in the plant growth-promoting bacterium *Pseudomonas fluorescens* SBW25. *Environ Microbiol* 2007;9:3046–3064. [PubMed: 17991033]
- Jordan, MI. Cambridge, MA: MIT Press; 1999. *Learning in Graphical Models*.
- Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology* 2006;7:198–210. [PubMed: 16496022]
- Kahlem P, Birney E. Dry work in a wet world: computation in systems biology. *Mol Syst Biol* 2006;2:40. [PubMed: 16820781]
- Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics* 2002;18:S225–S232. [PubMed: 12169551]

- Kay, SM. Englewood Cliffs, NJ: Prentice-Hall PTR; 1993. Fundamentals of Statistical Signal Processing.
- Kitano H. Systems biology: A brief overview. *Science* 2002;295:1662–1664. [PubMed: 11872829]
- Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 2005;3:e134. [PubMed: 15799710]
- Kschischang FR. Codes defined on graphs. *IEEE Commun. Mag* 2003;41:118–125.
- Lahdesmaki H, Hautaniemi S, Shmulevich I, Yli-Harja O. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. *Signal Processing* 2006;86:814–834. [PubMed: 17415411]
- Liu Y, Li J, Sam L, Goh C-S, Gerstein M, Lussier YA. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS computational biology* 2006;2:1419–1435.
- Loh KD, Gyaneshwar P, Markenscoff Papadimitriou E, Fong R, Kim KS, Parales R, Zhou Z, Inwood W, Kustu S. A previously undescribed pathway for pyrimidine catabolism. *Proc Natl Acad Sci U S A* 2006;103:5114–5119. [PubMed: 16540542]
- Mols M, de Been M, Zwietering MH, Moezelaar R, Abee T. Metabolic capacity of *Bacillus cereus* strains ATCC 14579 and ATCC 10987 interlinked with comparative genomics. *Environ Microbiol* 2007;9:2933–2944. [PubMed: 17991024]
- Müller, OF. *Vermivm terrestrium et fluviatilium, seu animalium infusoriorum, helminthicorum et testaceorum, non marinorum, succincta historia*. Havniæ, Lipsiæ: apud Heineck et Faber; 1773.
- Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* 2007;23:2322–2330. [PubMed: 17599939]
- Oh Y-K, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *The Journal of Biological Chemistry* 2007;282:28791–28799. [PubMed: 17573341]
- Ouchi, F. The World Bank. Report no. 3201. 2004. A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis.
- Pavord, A. New York: Bloomsbury; 2005. *The Naming of Names : The Search for Order in the World of Plants*.
- Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 2001;17:S215–S224. [PubMed: 11473012]
- Perkins AE, Nicholson WL. Uncovering new metabolic capabilities of *Bacillus subtilis* using phenotype profiling of rifampin-resistant *rpoB* mutants. *J Bacteriol* 2008;190:807–814. [PubMed: 17644585]
- Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886–897. [PubMed: 15494745]
- Price ND, Papin JA, Schilling CH, Palsson BO. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol* 2003;21:162–169. [PubMed: 12679064]
- Reed JL, Palsson BO. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J Bacteriol* 2003;185:2692–2699. [PubMed: 12700248]
- Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 2006;103:17480–17484. [PubMed: 17088549]
- Roweis ST, Ghahramani Z. A unifying review of linear Gaussian models. *Neural Computation* 1999;11:305–345. [PubMed: 9950734]
- Savageau, MA. Reading, MA: Addison-Wesley Pub. Co.; 1976. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*.
- Schafer, J.; Strimmer, K. Learning large-scale graphical Gaussian models from genomic data. In: Mendes, JFF.; Dorogovtsev, SN.; Povolotsky, A.; Abreu, FC.; Oliveira, JG., editors. *Science of Complex Networks: From Biology to the Internet and WWW*; CNET 2004. Aveiro, Portugal: American Institute of Physics; 2004. p. 320
- Searls DB. Data integration: challenges for drug discovery. *Nat Rev Drug Discov* 2005;4:45–58. [PubMed: 15688072]

- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–2504. [PubMed: 14597658]
- Shi J, Romero PR, Schoolnik GK, Spormann AM, Karp PD. Evidence supporting predicted metabolic pathways for *Vibrio cholerae*: gene expression data and clinical tests. *Nucleic Acids Res* 2006;34:2438–2444. [PubMed: 16682451]
- Shmulevich I, Dougherty ER, Zhang W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE* 2002a;90:1778–1792.
- Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 2002b;18:261–274. [PubMed: 11847074]
- Smyth P. Belief networks, hidden Markov models, and Markov random fields: A unifying view. *Pattern Recognition Letters* 1997;18:1261–1268.
- Sneath, PHA.; Sokal, RR. *The Principles and Practice of Numerical Classification*. San Francisco, CA: W. H. Freeman; 1973. *Numerical Taxonomy*.
- Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* 2004;101:2981–2986. [PubMed: 14973197]
- Tanay A, Steinfeld I, Kupiec M, Shamir R. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol Syst Biol* 2005;1 2005 0002.
- Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* 2002;18:287–297. [PubMed: 11847076]
- Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci U S A* 2003;100:8348–8353. [PubMed: 12826619]
- Uruburu F. History and services of culture collections. *Int. Microbiol* 2003;6:101–103. [PubMed: 12811589]
- Voit, EO. New York: Cambridge University Press; 2000. *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*.
- Wiener, N. New York: J. Wiley; 1948. *Cybernetics; or, Control and Communication in the Animal and the Machine*.
- Wolkenhauer O. Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform* 2001;2:258–270. [PubMed: 11589586]
- Woolfolk CA, Stadtman ER. Regulation of glutamine synthetase. 3. Cumulative feedback inhibition of glutamine synthetase from *Escherichia coli*. *Arch Biochem Biophys* 1967;118:736–755. [PubMed: 4860415]
- Xia Y, Yu HY, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao HY, Gerstein M. Analyzing cellular biochemistry in terms of molecular networks. *Annual Review of Biochemistry* 2004;73:1051–1087.

Figure 1. Biological Resource Centers preserve thousands of species and strains of microorganisms. This shows liquid nitrogen tanks that are used to store thousands of vials of cell lines and microorganisms at -170°C . Picture courtesy ATCC.

Figure 2.

Rank-abundance curves based on data from the NCBI Taxonomy. Panel (a) shows the distribution of gene records linked to the properly formed (Latin binomial) names of Bacteria and Archaea. Panel (b) shows the distribution of publications linked by PubMed IDs to the same names. Some of the most highly linked names are shown.

Figure 3.

An outline of the iterative process of building network models. Panel (a) encloses the initial data harvesting and preparation steps. Primary data are unprocessed output from sequencing, expression, phenotypic, and other experiments. Secondary data is the output of annotation and analysis of primary data. These data are stored in a database. Panel (b) encloses the data integration and network inference steps. Within a probabilistic graphical framework these two steps proceed in tandem and provide feedback to each other. Panel (c) encloses the network validation step. Both the network models and validation feedback are captured in order to refine the network inference. The gray lines represent the iterative cycle of network modeling; the dotted gray lines represent feedback loops.

Figure 4.

A protein-protein interaction map for *V. cholerae* based on data from the STRING database and visualized using Cytoscape (Shannon et al. 2003). The circles represent 3755 proteins, colored according to their functional class where known. Protein interactions at or above the 90% confidence level are shown as lines between two proteins. In this map, proteins that are encoded by a given locus appear only once. The proteins comprising the N-acetyl-D-glucosamine degradation, glycolysis and mixed acid fermentation pathways are the large aligned nodes on the right hand side of the figure.

Figure 5.

A metabolic map for *Vibrio cholerae* (Shi et al. 2006), produced by Pathway Tools. The complete map is on the left hand side of the figure. It contains many more pathways than actually occur in *V. cholerae*, but the other data used in network inference will constrain the pathways in the final model. In this representation, the pathways are represented discretely, meaning that the same enzymes can appear more than once in the map. This preserves the traditional textbook representation of the pathways. The circle on the right hand side shows that part of the map that includes the N-acetyl-D-glucosamine degradation pathway, which is traced by the green arrows. Enzymes are shown in red, substrates not in this pathway are in black, and substrates that are in the pathway are shown in green.

Figure 6.

A representation of an SML (systems biology markup language) model of all metabolic reactions in *V. cholerae* (Shi et al. 2006). The reactions of the N-acetyl-D-glucosamine degradation pathway and the Glycolysis I pathway are filled with color and shown on the right hand side of the figure. Round circles represent reactions (pathway reactions are red), while the diamonds represent species such as compounds (blue), regulators, and proteins (yellow) that participate in the reaction. The lines connecting the reactions and species represent reaction-modifier relationships (black) reaction-product relationships (green) and reaction-reactant relationships (red). This visualization, realized using Cytoscape, conveys the complexity of relationships in a metabolic network.

Table 1
Data sources, software tools and other resources

The table lists the resources and databases mentioned in the text. More comprehensive lists can be found in the annual database and web resources issues of *Nucleic Acids Research*.

Resource Name	Comment	URL *
GOLD	On line tracking of genome sequencing projects.	http://www.genomesonline.org/
PubMed	The main on-line literature database for biology.	http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed
GO	The gene ontology project provides a controlled, hierarchical vocabulary for the description of biological phenomena.	http://www.geneontology.org/
GenBank	The primary gene sequence database in the US.	http://www.ncbi.nlm.nih.gov/Genbank/index.html
JCVI CMR	The Comprehensive Microbial Resource site offer tools and data for the analysis of microbial genomes.	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
IMG	The Integrated Microbial Genomes system is a database of genetic, functional, and genomic data centered on genomic sequences.	http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
UniProt	An online source for data on proteins.	http://www.pir.uniprot.org/
InterPro	A database resource for identifying conserved elements in new proteins.	http://www.ebi.ac.uk/interpro/
pfam	A database of protein families, classified on the basis of functionally conserved sequence regions known as domains.	http://pfam.sanger.ac.uk/
STRING	Search tool for the recovery of interacting genes/proteins contains information on the interactions between thousands of proteins..	http://string.embl.de/
MetaCyc	A database of over 900 biochemical pathways. Reaction, kinetics, literature citations and other information are included.	http://metacyc.org/
BioCyc	A collection of 371 pathway genome databases.	http://biocyc.org/
KEGG	The Kyoto encyclopedia of genes and genomes.	http://www.genome.jp/kegg/
Reactome	A curated database of biological pathways and reactions.	http://www.reactome.org/
PlasmoDB	An online source for omics data relevant to the malaria parasite and its relatives.	http://plasmodb.org/plasmo/
Pointillist	Software that supplies various methods for integration of data.	http://magnet.systemsbiology.net/software/Pointillist/

* Note: checked on 2 May 2008