Enterprise Biology Software: I. Research (2001)

ROBERT P. BOLENDER

Enterprise Biology Software Project, P.O. Box 303, Medina, WA 98039-303 USA
http://enterprisebiology.com

Summary

An enterprise approach to biology was developed by combining biology, mathematics, and technology. The primary goal of the project was to develop a mathematical platform for biology and to use it as a discovery tool. New data derived from the biology literature suggested that principles identified long ago for physics and chemistry can apply to biology as well. Moreover, a new principle appeared that defined simple to complex connections within and across all levels of the biological hierarchy. These connections, which were often modular, could be assembled into networks and used to predict events – near and far. Taken together, these results suggest that complex events in biology can be a function of such networks. If this pattern continues to appear, then mapping networks and their interactions may provide a new way of exploring biology. The *Enterprise Biology Software* includes a short story, three courses (biology, mathematics/stereology, technology), three databases (courses, literature, tutorial), electronic data entry forms for research publications, and supporting material. It was written specifically for undergraduate biology students.

Introduction

The decision to map the human genome thrust biology abruptly into the Information Age. Today, genome data are routinely stored, accessed, and explored electronically. This is just the beginning of a new biology. The next task for biologists will be to translate the genome data into gene function, which ultimately involves all the structures and functions of biology. The *Enterprise Biology Software* explores ways of addressing this task, using guidelines developed at the *Matrix of Biological Knowledge Workshop* (Santa Fe Institute, 1987) and in the report: *Models for Biomedical Research, A New Perspective* (National Science Foundation, 1985). It extends the work of an earlier project (Bolender and Bluhm, 1992).

In business, the goal of an enterprise approach is to optimize profit. It identifies the best ways of using information to run a corporation most successfully. What would happen if we took this remarkable technology and applied it to a basic science such as biology? For example, can an enterprise approach to biology create opportunities for solving problems today that might otherwise be reserved for the future? The *Enterprise Biology Software (version 1.0)* attempted to answer this question by inventing the future – step by step.

The first step consisted of selecting a problem in biology sufficiently difficult to make the effort worthwhile. A list of the "most challenging" problems might include things like explaining gene function, development, or the brain – or understanding how a cell works or how biology manages a massive collection of parts that change relentlessly. Given such a list, the next step consisted of concluding that all such problems were subsets of a larger one - the one that qualified as being "sufficiently difficult." Remember biology has already solved all such problems and our goal here was merely to find ways of accessing existing solutions by applying mathematics and technology.

Consider this. Extending the frontiers of science depends – inescapably – on mathematics. Unfortunately, biology is largely a descriptive (soft) science advancing without the benefit of a mathematical platform. This deficit often leaves us ill equipped to attack even the simplest of complex problems. In contrast, physics and chemistry have robust mathematical platforms and have advanced our knowledge of the physical world for centuries. Perhaps, biology can never be a great science until it too has a mathematical platform – one not only equal to but more powerful than the ones enjoyed by the physical sciences. This need is unmistakable. Biology – by default – starts with all the complexity of physics and chemistry and then adds generous portions of its own.

A sufficiently difficult problem might therefore consist of building a mathematical platform for biology, of demonstrating that it can be used for discovery in ways typically reserved for physics and chemistry, and of creating resources to make this platform operational today. In short, an enterprise approach should allow us to access and use the mathematical principles of biology - routinely. The purpose of this paper is to

introduce the *Enterprise Biology Software* as an enabling technology and to include representative examples of what it can do. See the software package for additional examples and details.

Before beginning, however, a word of caution is in order. Our ability to explore biology as a mathematical science depends on a process of understanding that begins with hands on experience. Reading this paper – or perhaps any other paper - will not provide such an experience. Instead, the reader must be willing to try something new. Understanding within the framework of enterprise biology requires a transition from passive to active – it encourages the reader to move from the static pages of a journal or textbook to the interactive screens of a software package.

Methods

Procedures were developed for identifying the best research data, organizing these data into a mathematical platform, finding generalizations, and making predictions.

Working Rules

Seven rules served as guidelines for developing a mathematical platform for biology. The first purpose of the platform was to provide access to data that could be generalized.

Rule 1 - Unified Data: Find common roots in experimental biology.

Rule 2 - Quantitative Structure: Relate data to three-dimensional space.

Rule 3 - Structural Hierarchies: Link data across biology.

Rule 4 - Unbiased Sampling: Collect data without bias.

Rule 5 - Balanced Equations: Apply mathematical principles to biology.

Rule 6 - Critical Data: Detect biological changes unambiguously.

Rule 7 - Relational Databases: Organize the biology literature.

Working Models

The seven rules were translated into three models, which when combined defined the enterprise approach to biology.

Model 1 - Qualitative Hierarchy: Identify structures and organize them by defining their positions within and across hierarchical levels. The purpose of the model was to standardize terminology, arrange structures according to a standard pattern, and to serve as a guide for entering published data into the biology literature database. Applications included the brain and hierarchy browsers.

Model 2 - Quantitative Hierarchy: Define an organism as a three-dimensional space containing a great number of nested volume compartments. The model allowed a question in biology to be structured as a balanced equation, identified the variables needed to detect change unambiguously, and supported data summaries across an entire literature. Here the hierarchy - defined explicitly with equations - was explored with simulators, tutorials, and real data.

Model 3 - Relational Database: Translate the first two models into a set of tables and relationships. The model allowed data to be entered, filtered, sorted, explored, standardized, and displayed in ways determined by the user.

Strategy

The overall plan of the project included two design strategies. First, insure long term flexibility of the mathematical platform by starting with nothing and ending up with practically everything. Second, manage complexity by going from simple to complex to simple… - and by going from complex to simple to complex...

Biology is a product of nature and as such obeys the laws of nature. In contrast to physics and chemistry, however, the relationship of biology to these laws remains largely a mystery because we still do not know how to derive biology mathematically from first principles.

Since biology is derived from the physical world of physics and chemistry, it must be subject to the laws governing matter – along with an undetermined amount of law not yet discovered. By this definition, biology encapsulates the complexity of physics and chemistry along with all of its own. Although abundantly present, laws of nature may be deeply hidden by this complexity. If true, then a strategy for gaining access to these fundamental principles of biology might consist of merely peeling away this complexity layer by layer.

To test such an idea, a list of layers was prepared. Since organization is the first step toward understanding complexity, the peeling process consisted of organizing and simplifying each layer of complexity. Understanding one layer of complexity led to understanding the next layer and then the next…

Layers of Complexity

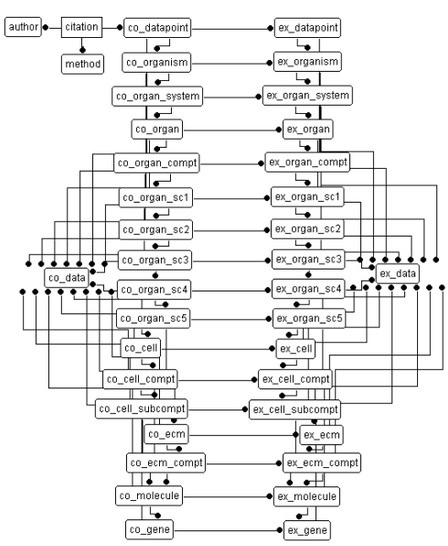
Layer	Understanding
Elements	Derive all data from a common source.
Organization	Arrange biology as a hierarchy of size - structures within structures.
Change	Detect change unambiguously.
Relational Model	Translate experimental biology into a database model.
Papers	Find the best ones for this project.
Standardization	Express all data as numbers with structural locations.
Connections	Connect data to search for new patterns and principles.

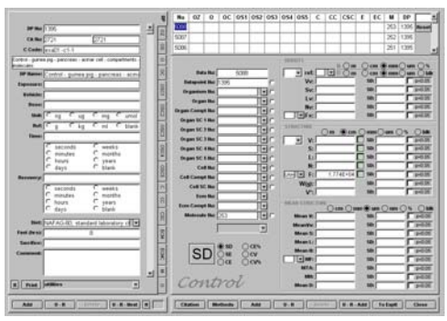
Since a detailed description of moving through these layers was included with the software (Short Story, Chapter 2), it will not be duplicated here. Moreover, understanding the fine points of the process benefits importantly from using simulators, viewing research data, and taking tutorials.

Biology Literature Database

The goal of standardization was to take information from highly heterogeneous sources (research publications) in such a way that thousands of wholly unconnected experiments became connected – mathematically. Here, standardization became a product of the data entry process. The primary requirement of the literature database model was that it should accommodate all or most biological data – including both structure and function.

Relational Database Model: In short, the database design included a structural hierarchy consisting of sixteen compartments (defined), twelve structural data types (defined), and three functional data types (user-definable). It also included tables for author, citation, and method.





As shown in figure at the left, the literature database used two structural hierarchies – one for control data points and the other for experimental. The hierarchies were connected to one another and to either a

control or experimental data table (co_data, ex_data). The figure represents a logical database model wherein entities (boxes) have relationships (lines). In a physical database model, the entities represent database tables containing columns and rows. Such tables were presented to the viewer either as data entry forms (shown at the right) or as literature browsers.

Research data were taken from a refereed publication and entered into the database one data point (time point) at a time. The process of data entry consisted of first building a structural hierarchy for the data point and then mapping numerical data to it. Structural data included volume, surface, length, and number. Both structural and functional data could be related to a unit of volume (density), to a structure, or to an average structure. It took about four hours to move the data from one publication into the database. This included reading the paper, converting data to a numerical format, keying in the data, and checking for errors. Data entry tools were used throughout to speed the process.

Separate data entry screens were written for a production environment and for the occasional user. These screens, for example, will allow a research laboratory to convert their data into an electronic format and then use a variety of analysis screens to view their results embedded in the larger literature database. In effect, the database serves as an enabling technology for investigators interested in comparing results, looking for patterns in their data, or publishing their data electronically.

Selecting Papers for Data Entry: The biology literature includes research papers numbering in the millions. While moving all these data into a literature database is theoretically possible, the goal here was to be selective. Papers of interest included those with hierarchical data because the discovery strategy depended largely on making new connections. A search of the biology literature for methods currently being used to collect such data identified two main candidates - morphometry and stereology.

Morphometry refers to the measurement of form, whereas stereology identifies a specific statistical approach for estimating four of our original elements (points, length, surface, volume). Stereology seemed the better choice because it fitted quite well with the organizational framework and because it could operate throughout n-dimensional space. It also offered unbiased methods for collecting data and could detect change unambiguously (Baddeley et al., 1986; Gokhale, 1990; Bolender et al., 1993).

Citations were collected from ***Medline,*** a bibliographic service run by the National Library of Medicine, for the years 1965 to 1998. A keyword search using stereolog$ generated a list of all papers with stereology, stereologic, and stereological in the title or abstract. If - after reading the abstract - the paper appeared to meet the requirements for data entry, a copy was ordered from our university copy service. This first cut gave more than three thousand papers. Each paper was read and graded with a checklist. If one of the following questions was answered yes, then the data of that paper were entered into the literature database: (1) Was change detected reliably? (2) Were unbiased methods used to collect data? (3) Could the results be recalculated to satisfy the requirement of question 1? Roughly, five hundred papers were eventually selected for data entry.

Data

By applying strict standards for data entry, it was thought that the resulting literature database would have the best chance of supporting a mathematical platform. The first step of the data analysis included looking for patterns in the data that could be generalized.

Two all-purpose views of the biology literature were used. The first view included screens that displayed standardized data for citations, methods, and data (control and experimental). This group included graphs, percentage change, phenotype, and SQL sort. The second view included screens that generated new data from the old – as a way of finding generalizations.

Standardized Data

Biology Literature: Each paper in the database was summarized – in standard form – by citation, author, method, hierarchy, and research data.

Graphs: Graphical summaries were included for citation, author, method, country, organ system, and organ.

Percentage change: A percentage change was defined as an experimental value divided by the control and multiplied by 100% (% Change = (Experimental / Control) x 100%). In these screens, data were highlighted according to increase (red>100%), decrease (blue<100%), and no change (green=100%).

Phenotype: The phenotype screen included percentage change data displayed across fourteen levels of the hierarchy – from organism to molecule. The screen was designed to show the pattern of change in one or

all hierarchical compartments of a given data point – for all data points in the database. It was originally thought that it might serve as a reverse-engineering tool.

SQL Sort: The citation and methods tables of the literature database were used as an example of how complex SQL (Structured Query Language) queries could be generated automatically – in response to selections made by the user.

Generalized Data

Connection Maps: Since an analysis of the first view (standardized data) yielded few patterns that could be generalized, the second view generated all possible combinations of the research data – paper by paper. These results were expressed as ratios and in turn plotted as regressions (generalized). The regression equations were grouped according to the connection map type. For details, see the short story in the software.

The connection types included proportionality constants: one structure vs. one structure at several time points - at one level (Connection Map – Type 1), many structures at one time point vs. many structures at another time point - at one level (Connection Map – Type 2), and many structures at one time point vs. many structures at another time point – at several levels (Connection Map – Type 3). The general equation of the linear regression curve (y=mx+a) defined the relationship between y and x wherein k=y/x, k being the proportionality constant. These analyses detected numerous quantitative relationships between and among structures, often with coefficients of determination (r^2) approaching 1.0. When overlap occurred between pairs of structures, connection maps could be drawn (open, closed, and branched). This method allowed generalization, but only for specific experimental settings.

To generalize data across biology, structures were plotted against structures – using all possible permutations of the data (Connection Map – Type 4). Here, however, regressions were plotted as power curves ($y=bx^a$). In turn, the power equations - with r^2 approaching 1.0 - were fitted into data replicators where they served as prediction algorithms. Case studies (*.doc) and calculation scratch files (*.xls) were included with the software to illustrate the methods of analysis.

Management Tools

Tools for maintaining the courses and databases were included in the software appendix.

Development Tools

Programs: Programs and databases were written with PowerBuilder Enterprise 6.5 (Sybase, Inc., Emeryville, CA). Slide shows and individual lecture slides (wmf files) were written with PowerPoint (Microsoft, Redmond, WA), regression data were calculated with Microsoft Excel, and tutorials were written with Microsoft Word 2000.

Database: The Sybase SQL Anywhere 5.0 (Sybase, Inc., Emeryville, CA) database engine was used for development; a runtime version of the engine was distributed with the software package. Copies of the courses, literature, and tutorial databases were also distributed.

Database Design: The relational models of the databases were developed using ERwin (Computer Associates, CA).

Images: Light and electron micrographs were taken by the author and labeled in Photoshop 5.5 (Adobe Systems Inc., San Jose, CA). Three-dimensional images, which were licensed from Viewpoint Datalabs International Inc., Orem, UT), were positioned, labeled and rendered in 3D Studio Max R2 (Kinetix /Autodesk, Inc.), San Francisco, CA). Line drawings were licensed from Williams and Wilkens, Baltimore, MD) and others from CorelDRAW (Ottawa, Ontario). Thumbs-Plus (Cerious Software, Inc., Charlotte, NC) was used to maintain the image collection.

Support Files: Help files were written with RoboHELP Classic (Blue Sky Software Corporation, La Jolla, CA) and installation programs with InstallShield Express 3.03 (InstallShield Software Corporation, Schumburg, IL).

Web Site: The Enterprise Biology Software site was written with Microsoft FrontPage 2002.

Results

Software Package

The ***Enterprise Biology Software*** contained 1.2 GB of programs, files, and databases. This included 352 programs, 528 files, 1,500 images, 1,200 short answer questions, and 3 relational databases. After installing the software, the installation program automatically sets up a client-server facility on the target PC computer.

Biology Literature Database

A relational database for the biology literature was designed and implemented, using a rule-based approach. Thus far, the database has accommodated all or most types of structure and function data encountered during the data entry process. Taken together, the database and associated interface screens demonstrate the feasibility of standardizing the biology literature and of using the literature to assemble a mathematical platform for biology.

Change

Change in biology was explored extensively, using simulators and data taken from the literature. The general pattern to emerge from this effort suggested that a change typically involves many parts of a structure that can change at different rates and in different places. Although rules and methods could be applied to detect change unambiguously at a given time and place, change was seen as a complex event and interpreted either as a continuous event or as two equilibrium states separated by a transition. Moreover, the results suggested that change could be expressed as a function of structural networks connected by rule

Generalization

Typically, research data published in the original papers were not used by the authors to look for generalizations. Instead, the data were collected and interpreted primarily to answer specific questions within the boundaries of a specific experimental design. This meant that only a small portion of the published data of a study was actually being used. Here, however, where generalization was a primary goal, all permutations of the data were considered. The effort was rewarded in that an exhaustive analysis yielded widespread generalizations within and across publications. This finding suggests that structural data – arranged within an enterprise framework - opens a small but distinct window into the mathematical core of biology.

The new data were summarized as regression equations and in turn used to construct connection maps. The simplest connection map consisted of a relationship between two structures or functions, defined by a regression equation with an r^2 approaching 1.0. However, these simple maps were often found to be part of a larger map – occurring within or across levels of the hierarchy. The maps displayed connection sets that were open, closed, or branched. Examples of these connections were included in chapter 2 of the short story.

Taken together, the connection maps suggest that the structural organization of biology consists of networks contained within networks. Case studies were included with the software to illustrate how connection maps could be assembled and interpreted.

Prediction

Data replicators were built from the connection maps and used to predict structure. The first replicator (one from one) included connections between two structures, being defined by a single power equation ($b_1X_1^{a1} = Y_1$). When a structure was selected from a list, all the other structures with connections to it were displayed. The software predicts a value for the connected structure after a seed value is entered. For

example, when 1000 was entered for surface area of the rer, the predicted surface area of the nuclear membrane was 34.8. (N.B., this prediction was based on data taken from several independent publications and had an r^2 of 0.9810).

The second replicator (many from one) included connections among several structures identified by overlapping similarities. Here the algorithm included several overlapping power equations.

Equation 1: $b_1X_1^{a1} = Y_1$

Equation 2: $b_2X_2^{a2} = Y_2$, where $X_2 = Y_1$

Equation 3: $b_3X_3^{a3} = Y_3$, where $X_3 = X_2$

Equation n: $b_nX_n^{an} = Y_n$, where $X_n = X_{n-1}$

The many from one replicator worked similarly. When a structure was selected from a list, a screen appeared containing all the connected structures. After a seed value was entered, predictions were returned for all the connected structures. For example, a brain with 1000 neurons in the parietal lobe would predict 1,420,000 neurons in the temporal lobe, 3,243,000 in the frontal, and 5,662,000 in the occipital lobe. As a control, the replicator algorithm also predicted the original seed value. For the example above, 1,028 neurons would be predicted for the parietal lobe instead of the expected 1000, an error of 2.8%. Here, predictions were based on data taken from one or several publications.

Case Studies

Case studies were included to illustrate potential applications of the connections model to experimental biology. Taken together, they suggest a mathematical view of biology similar to the one enjoyed by the physical sciences.

Appendix

The appendix included tools for maintaining the courses and databases. It provides a realistic view of what it takes to build and support an enterprise approach to biology - in an academic setting.

Discussion

This project began by imagining what biology might be like in the future and then proceeded to try out that future now. The ***Enterprise Biology Software*** was the result. It included a collection of programs, files, simulators, and databases that together offered new ways of exploring biology.

A primary goal of any enterprise approach is to build an information system that optimizes outcomes. In business, profit is the usual goal; here, it was learning and discovery. Early in the development process, the physical sciences became the model to emulate. These “hard” sciences have advanced learning and discovery for centuries because they can operate from a mathematical platform. Such a platform has allowed scientists to find generalizations in experimental data that eventually pointed to the principles and laws of nature. Here a similar mathematical platform was imagined for biology and a major challenge of the project was to assemble such a platform and try it out.

The release of the ***Enterprise Biology Software*** serves to demonstrate that such a platform could be built and used as imagined. Moreover, when biological data were viewed from a mathematical platform they behaved in ways surprisingly similar to those of the physical sciences. In biology, however, data well suited to finding generalizations are not being published routinely. In most cases, new data had to be generated from the original studies before generalizations could be found. Historically, the standard experimental model used in most biological studies tracks structural or functional changes in response to an event or treatment.

In contrast, the view from the mathematical platform suggests that change is – at least in part – a function of the connections between structures. This observation may become important in the future because identifying such connections may be one of the critical steps in unraveling complexity in biology. Without a better understanding of change, we may end up looking for the right answers in the wrong places.

A Mathematical Platform for Biology

In physics and chemistry, experimental data are fitted to curves that can be expressed as equations. This represents the standard empirical approach to discovery. The platform for biology applied a similar technique. Relationships between two structures were catalogued mathematically as proportionality constants and as power functions. In addition, relationships among several structures were reported as connection maps. These interactions were interpreted as generalizations either at the level of an individual experiment or across the biology literature. In turn, generalizations were used to predict structure with a reliability based on empirical equations with r^2s approaching 1.0. In short, the mathematical platform for biology successfully duplicated two properties of the physical sciences – generalization and prediction.

A curious pattern emerged from the connection maps. In effect, they defined how the parts of biology fit together – as networks within networks. This suggests a new level of order in biology - the mechanism of which remains unknown. One can imagine that such order occurs spontaneously or by design. Perhaps this is an example of where a previously undetected law of nature is finding expression within the complexity of biology. If, for example, the networks turn out to be optimized, then one might surmise that biology naturally tends toward a best (optimal) solution. Such ideas are already being tested (Waldrop, 1993; Kauffman, 1995; Brown et al., 2000; Alexander and Enquist, 2000; Schreiner et al., 2000). Curiously, these new connections and networks appeared only after the effect of time - as a variable - was removed from the data. This underscores the importance of simplification in the discovery process – a well-known tradition in the physical sciences.

Recall that the mathematical model used to design the literature database was hierarchical, consisting of compartments contained within compartments. In the future, however, databases designed for a networks contained within networks model may offer a more direct way of connecting higher-level structures to molecules and genes.

Laws of Nature

In theory, biology is a mathematical science. In practice, however, the prognosis remains guarded because biology is being held in "irons." It is effectively prevented from gaining a quantitative status because the current library model – books on shelves – isolates research data and this isolation inhibits the formation of connections. Connections become a critical path when they lead to the mathematical core of biology.

Data City was written as a short story to give the reader a brief glimpse of a biology freed from this constraint. It showed that mathematical generalization in biology becomes apparent only when we have ready access to data from many papers – simultaneously – in standard form. Since mathematics is the primary route to discovery in science, the short story looked for ways of exploring biology mathematically. It assembled a mathematics-based strategy, found the best data, organized them, connected them, and stored them electronically. In turn, the data were used to look for generalizations, principles, candidate theories, and new discovery models.

Principle of Scaling: Power laws occur throughout the natural world. They can be found, for example, in astronomy, biology, economics, history, meteorology, music, and physics. The slope of a scaling law gives a precise view of the rules operating in a system. Similar slopes suggest similar underlying rules and different slopes different ones. The power law is expressed as $Y = bX^a$, where Y is the dependent variable, b the y intercept, X the independent variable, and a the slope.

An excellent book edited by J.H. Brown and G.B. West entitled "Scaling in Biology" provides compelling evidence that biology obeys scaling laws ranging all the way from molecules to ecosystems. More importantly, several authors demonstrated that biological data could be predicted directly from first principles (Brown et al., 2000, West et al., 2000; Alexander, 2000; Schreiner et al., 2000) – without the use of empirical data. Moreover, a growing body of evidence strongly suggests that biology behaves as an optimizing system.

Principle of Stochiometry: Stochiometry is the calculation of the quantities of chemical elements or compounds involved in chemical reactions. It identifies the exact proportions required. Although the results showed that structures in biology obey the scaling principle across hierarchical levels and animals, order was also found within levels of the same animals – in individual publications. In case study 2, for example, an equation was written for the human cerebellum – as a function of cell connectivity. In effect, the equation identified the exact proportions of cells ("elements") required to make a cerebellum ("reaction"). Indeed, the equation suggested that a law analogous to the one governing chemical reactions

might be operating at the levels of organelles and cells. Alternatively, the equation – and the cerebellum - might be an emergent property of the underlying chemical stochiometry. Here, emergence refers to local interactions generating large-scale effects that do not exist in the individual parts.

Candidate Theories

The **theory of biological connections** states that structure is ordered mathematically within and across all levels of the biological hierarchy. This theory, which suggests widespread structural order throughout biology, gathers support from the literature database as more than 1000 regression equations with r^2 approaching 1.0. These data are included in the software.

The **theory of biological change** states that change is a process occurring simultaneously at different rates within and across levels of the biological hierarchy. This theory suggests that an experimentally induced change exhibits local to global effects. The supporting evidence includes more than 800 regression equations with r^2 approaching 1.0.

Change

If the many parts of biology are connected by rule, then a change in any one part may trigger a global response of astonishing complexity. The study focused first on detecting change reliably and then used the new data to look for patterns of change.

Plots of experimental data identified multiple outcomes. Change was expressed either as parallel or non-parallel curves, which suggested equilibrium (parallel curves) or a transition toward equilibrium (non-parallel curves). Parallel curves indicated more or less of the same structures (equally), whereas non-parallel curves indicated more or less of similar structures (unequally). The point to make here is that change behaves as a complex event and its interpretation depends importantly on when and where it is observed. A task remaining is to look at change more closely by dissecting it into smaller and smaller parts – at both equilibrium and non-equilibrium states.

Dissecting biochemical change, for example, suggests that a cell can respond to a stimulus first by designing a new membrane and then by actively producing it in quantity (Bolender, 1981). Although equilibrium appeared in the biochemical composition of the membrane after one day, the amount of the membrane in a cell continued to increase throughout five days of treatment. Curiously, the method used for estimating the concentration of an enzyme in a membrane relied on the simultaneous solution of linear equations. Recall that such a solution is used routinely for finding the best outcome.

Please note that data from controls – but not experimentals - were used to calculate connection maps across the biology literature. Therefore, the generalizations reported as type four connection maps refer exclusively to control data.

Stereology

When looking for hierarchical patterns in biology, the best papers offered the best chance of finding the most reliable results. Stereology provided the best source of such data, because the sampling methods (Baddeley et al., 1986; Gundersen, 1986; Gokhale, 1990) provided unbiased estimates for all the many different parts of biology. Indeed, without the guiding rules of stereology (Weibel, 1979) the task of constructing a mathematical platform with empirical data could not have been attempted.

In reality, however, stereological estimates can be biased when inconsistency occurs between theory and practice. Well-known sources of bias in stereology include such things as section thickness, overlapping structures, and shape. Moreover, methods of tissue preparation can influence the extent to which material is retained, lost, or changed (Bertram et al., 1986). By forming ratios of structures – as done in the analysis here – the influence of such bias was often minimized. However, any minimization of bias from data point to data point or from paper to paper could not be considered uniform. Even the best data may therefore carry an unknown bias.

The mathematics of stereology can extend across n-dimensional space. In physics, for example, a major path to discovery has consisted of breaking matter into smaller and smaller parts. In recent years, physicists have begun to study these parts in dimensions well beyond the third (Greene, 1999). In the future, however, access to these higher dimensions may also become essential for biologists.

Economics

The software included a cost-benefit analysis of the biology literature database – including the effect of generating new data from old. An important finding of the project was that only a small fraction of the information available in a publication was currently being used. By generating all possible outcomes of the results, many fold gains in productivity can occur routinely. This identified an important source of high quality data that can be used to expand our knowledge of structural networks substantially - at relatively low cost.

The inherent value of electronic data entry is that it leverages the original investment in biological research by generating new data from old. For example, doubling the amount of useful information in a research paper is roughly equal in value to the cost of generating a new paper. Set the unit cost of producing a new paper, and the potential value of even a small literature database - containing as few as 500 papers - becomes surprising. A biology literature database therefore offers an excellent strategy for increasing research productivity.

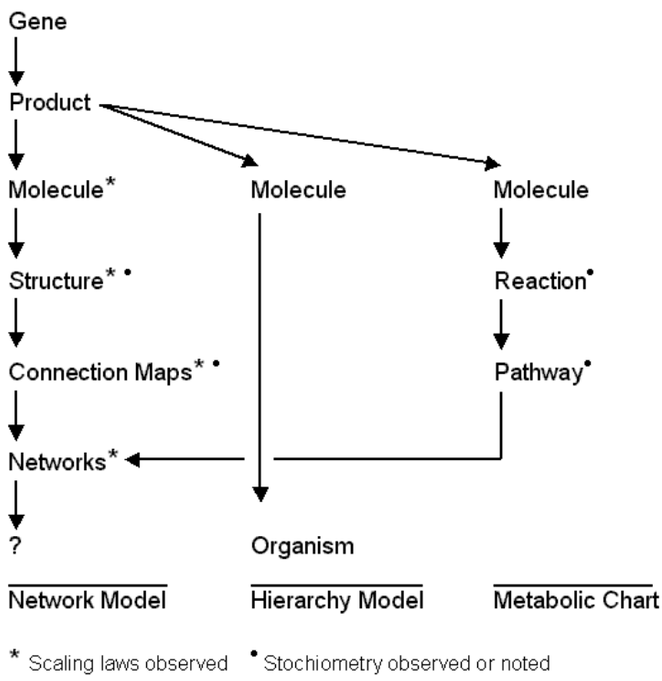
Research Models

Change the base of biology from descriptive to quantitative and the entire research environment changes. The design of the biology literature database reflected – by necessity – the direction of a research community largely interested in detecting change. When the database became operational, however, it became quickly apparent that the traditional model of change was not well suited to the task of finding generalizations. Unfolding complexity across biology required a new and more powerful model - one based on connections. The price of this model will be electronic data entry – across all disciplines.

A Working Model… An Emerging View…

The ***Enterprise Biology Software*** pursues new strategies for unraveling the complexity of biology by developing databases and research tools. Thus far, the software suggests that several things now seem possible; (1) The biology literature can be standardized and stored in a database; (2) Strategies can be identified for generating large quantities of new data from old; and (3) Scaling laws can be applied to decrease complexity of biological data and in turn reveal deeper patterns of mathematical order.

The figure below summarizes the progress of the research component of this software project. Scaling laws and stochiometry seem to account for all of the mathematical order observed thus far. The hierarchy model established an orderly framework for organizing structures and data, whereas its first progeny - the network model – points directly to familiar laws of nature. The metabolic chart is included with the prediction that it too will map as a network. The question mark suggests the next likely direction in the ongoing process of unraveling biological complexity.



Concluding Comment

The opening pages of the *Report of the Matrix of Biological Knowledge Workshop* (1987) carried the following quotes from the 1985 report *Models for Biomedical Research: A New Perspective*.

"We seem to be at a point in the history of biology where new generalizations and higher order biological laws are being approached but may be obscured by the simple mass of data."

[Biomatrix definition] ***"The complete database of published biological experiments, structured by the laws, empirical generalizations, and physical foundations of biology and connected by all the interspecific transfers of information."***

"The development of the matrix and the extraction of biological generalizations from it are going to require a new kind of scientist, a person familiar enough with the subject being studied to read the literature critically, yet expert enough in information science to be innovative in developing methods of classifications and search. This implies the development of a new kind of theory geared explicitly to biology with its particular theory structure. It will be tied to the use of computers, will be required to deal with the vast amount and complexity of information…"

Such statements inevitably guide one to the conclusion that biology in the future must be a science based in mathematics

References

Alexander, R. M. Enquist 2000 Hovering and jumping: Contrasting problems in scaling. In: Scaling in Biology, Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, New York, pp. 37-50.

Baddeley, A. J., H. J. G. Gundersen, and L. M. Cruz-Orive. 1986 Estimation of surface area from vertical sections. J. Microsc. 142:259-276.

Bertram, J. F., P. D. Sampson, and R. P. Bolender 1986 Influence of tissue composition on the final volume of rat liver blocks prepared for electron microscopy. J. Electron Microsc. Tech. 4: 303-314.

Bolender, R. P. 1981 Stereology: Applications to pharmacology. Ann. Rev. Pharmacol. Toxicol. 21: 549-573.

Bolender, R. P. and J. M. Bluhm 1992 Database literature review: A new tool for experimental biology. Mathl. Comput. Modelling, 16:11-35.

Bolender, R. P., D. M. Hyde, and R. T. DeHoff 1993 Lung morphometry of the lung: A new generation of tools and experiments for organ, tissue, cell, and molecular biology (Invited Methods Review: Am. J. Physiol. 265 (Lung Cell. Mol. Physiol. 9) L521-L548.

Bolender, R. P. 1993-1997 Human Biology Software © Robert P. Bolender and the University of Washington.

Brown J. H., G. B. West, and B. J. Enquist 2000 Scaling in Biology: Patterns and processes, causes and consequences. In: Scaling in Biology, Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, New York, pp. 1-24.

Committee on Models for Biomedical Research 1985 Models for Biomedical Research A new Perspective, National Research Council, National Academy Press, Washington, D.C.

Cruz-Orive, L. M. and E. R. Weibel 1990 Recent stereological methods for cell biology: a brief survey. Am. J. Physiol. 258 (Lung Cell. Mol. Physiol. 2) L148-L156.

Gokhale, A. M. 1990 Unbiased estimation of curve length in 3D using vertical slices. J. Microsc. 159: 133-141.

Greene, G. The Elegant Universe. 1999 Vintage Books/Random House, Inc., New York.

Gundersen, H. J. G. Stereology of arbitrary particles. 1986 A review of unbiased number and size estimators and the presentation of some new ones in memory of William R. Thompson. J. Microsc. 143: 3-45.

Kauffman, S. At Home in the Universe. 1995 Oxford University Press, New York.

Morowitz, H. J., and T. Smith. 1987 Report of the Matrix of Biological Knowledge Workshop. Santa Fe, NM. Santa Fe Institute.

Schreiner, W., R. Karch, F. Neumann, and M. Neumann. 2000 Constrained constructive optimization of arterial tree models. In: Scaling in Biology, Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, New York, pp. 145-165.

Walthrop, M. M. Complexity. 1992 Simon & Schuster, New York.

Weibel, E. R. Stereological Methods. Practical Methods for Biological Morphometry. 1979 Academic Press, London.

West G. B., J. H. Brown, and B. J. Enquist 2000 The origin of universal scaling laws in biology. In: Scaling in Biology, Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, New York, pp. 87-112.

Saturday, December 07, 2002
C:\EBiology\paper1.doc