Enterprise Biology Software: V. Research (2004)

ROBERT P. BOLENDER

Enterprise Biology Software Project, P. O. Box 303, Medina, WA 98039-0303, USA
http://enterprisebiology.com

Summary

The ***Enterprise Biology Software Project*** explores new approaches to discovery in the life sciences by applying mathematics and technology to the biology literature. This process includes (1) standardizing the literature by moving research data from the pages of journals into the tables of a relational database, (2) generating derived data libraries from the database, (3) searching these new libraries for mathematical patterns, and (4) distributing the databases, libraries, software, and observations freely to contributing authors - on a CD. The current release includes an updated stereology literature database, new libraries (***repertoire, analogy, drill-down, and ladder***), a progress report, and several new findings. In short, the libraries continue to uncover widespread patterns of mathematical order in biology. These patterns appear not as properties of individual structures, but rather as connections between structures. Connections, for example, define repertoires of equations that form networks. In the report, we consider how these equations might help us decipher the genetic regulatory networks.

Introduction

Background

The results of the **Enterprise Biology Software Project** suggest that the mathematical core of biology is well hidden by several factors, including the bias of our experimental methods, the fluctuation of biological systems between equilibrium and nonequilibrium states, and the confounding nature of change. Since the success of the project depends on accessing this mathematical core, all these issues had to be dealt with – locally and globally - in both control and experimental settings. This was accomplished by storing published research data in a relational database and then generating libraries of standardized data there from (Bolender, 2001-2004).

As the database grew, it became apparent that these new libraries (data pairs; design codes) could generate additional libraries consisting of equations. The equation libraries created a new opportunity for exploring biology as a mathematical puzzle, one that could be solved – one step at a time – by following the clues accompanying each new library. In effect, we now know how to translate the biology literature into collections of standardized equations. The challenge for the reader – and for the writer as well – is to find the clues and then use them to solve additional pieces of the biology puzzle.

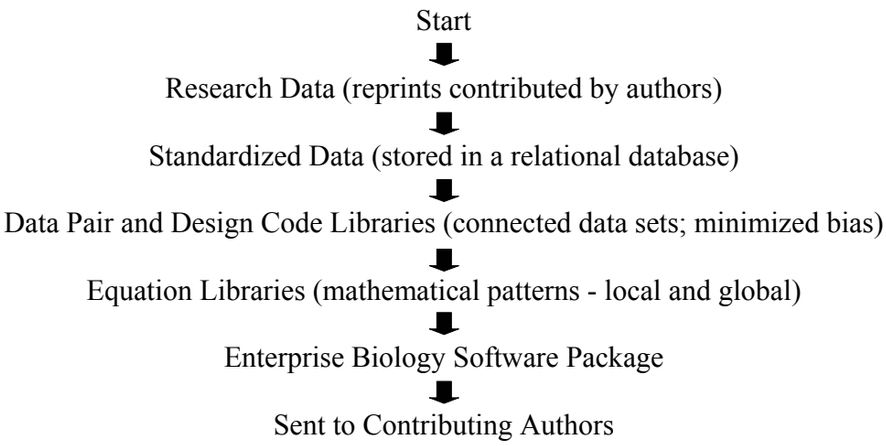
How do we find the clues? Last year, you may recall that the ladder equation library showed us that biological data – expressed first as data pairs and then as power equations – could be summarized by a single exponential equation. This suggested that all the parts of biology might be ordered by a single rule. The major clue to come from that exercise was that order could be found as power equations with $r^2 \geq 0.999$ by sorting the data pairs as ratios (y/x) – three or more rows at a time. In other words, we learned how to convert data pairs into equations - quickly. The next clue came from the observation that the rung equations grouped the data pairs at distinct levels, resembling quantum steps. A similar clue came from the design code library when it showed us that change behaved – quite unexpectedly - like a constant. Finally, the example of unfolding complexity by shifting our perspective from a zero to a one-dimensional platform provided the catalyzing clue.

Armed with so many clues, the next piece of the puzzle was in easy reach. Let us begin with the fourth clue, the one that recommended a shift in our perspective. The promise here was that changing the platform for viewing data can change what we see. Consider this. What would happen if instead of looking at biology thorough our eyes we looked at the same biology through the "eyes" of one of its parts? Although such a question seems a bit odd at first, it turns out to be a quite good one because it takes us to an interesting result. If we assume – for convenience - that any given part sees all those parts to which it is connected mathematically, then we can share the perspective of that part by simply writing the appropriate equations. The result becomes interesting to the larger community if the parts found to be connected mathematically turn out to have been produced by a similar genetic regulatory network. If true, then we have found a relatively simple way of reverse engineering these otherwise elusive networks. In short, the perspective clue - supported by the first three – produced a new library (repertoire), one that may be offering us our first glimpse of what organelles and cells may be "seeing." Along with our glimpse, however, comes a view of biology that – if confirmed – suggests a level of complexity for genetic regulation that may help to explain why genomes can do so much more than just code for proteins.

The other new libraries are more or less straightforward. One uses equations to suggest a strategy for connecting larger structures to molecules and genes (drill-down), whereas the other (analogy) employs equations to hunt for similarities in the biology literature - mathematically. Finally, a ladder equation library was added for experimental data.

Progress

Currently, the major activities of the ***Enterprise Biology Software Project*** consist of entering data and generating new libraries. Research data submitted by authors were added to the database and sent back as part of a technology package, as shown below.



First, data were moved from research papers into a relational database and standardized. For each paper, all the data at a given hierarchical level of size (1 to 16) were then used to form pairs of data, which juxtaposed control vs. control (data pairs) or control vs. experimental (design codes). These two basic data libraries were stored as database tables and used to generate the equations that populate the derived data libraries. Writing equations consisted of filtering and sorting the data, sending the results to an Excel worksheet, and plotting the x and y values as power equations ($y = bx^a$). Finding these equations was simplified by sorting the ratio y/x from low to high, and looking for an $r^2 = 0.999$ or better with three or more rows of data. The resulting graphs were stored in Excel files and included with the software upgrade. The libraries – old and new – are listed in Table 1.

Libraries: Libraries serve as discovery platforms (Table 1). They include one or more user interface screens, data, help files, and worked examples (e.g., Excel worksheets; case studies). In effect, each new library helps to solve another piece of the biology puzzle.

Table 1. Enterprise Biology Software Libraries.

Library	Data	Entries	Applications
Standardized Stereology Literature			
• Citation – search	original	12,853	Find references
• Citation – by paper – contl	original	1,024	Print paper – contl data
• Citation – by paper – contl + exptl	original	6,438	Print paper – contl + exptl data
• Methods – search SQL script	original	1,951	Find papers by methods
• Control Data	original	15,521	
• Experimental Data	original	9,677	
• Contl data – by data point	original	12,164	Find data by data point; level
• Contl+Exptl data – by data point	original	7,284	Find data by data point; level
• Percentage change data	derived	7,018	Find data by change; level
• Phenotype data	original	7,018	Find data across 14 levels
Connection Map			
• Type 1 (2str/2+points/1level/1paper)	derived	182	Find connections/minimize bias
• Type 2 (2+str/2+points/1level/1paper)	derived	81	Find connections/minimize bias
• Type 3(2+str/2+points/1+levels/1paper)	derived	323	Find connections/minimize bias
• Type 4 (data pairs)	derived	22,445	Find connections/minimize bias
Data Replicator			
• One from one (data from 1 paper)	derived	702	Predict data
• Many from one (data from 1+ papers)	derived	27	Predict data
Biological Algorithm			
• Connections upstream and down	derived	458	Predict organs and organisms
Data Pair			
Global (data from 1+ papers)	derived	112	Find connections/minimize bias
Design Code			
• Local (data from 1 paper)	derived	2398	Identify and predict change
• Global (data from1+ papers)	derived	58	Identify and predict change
Ladder Equation (Data Pairs)			
• Total data pairs	derived	25	Generalize structure in biology
• Organ	derived	19	Generalize structure by organ
• Cell	derived	19	Generalize structure by cell
• Organelle	derived	22	Generalize structure by organelle
New for 2004			
Repertoire (Data Pairs)			
• Organelles, inclusions, and cells	derived	771	Find connections; Make predictions
Ladder Equation (Design Codes)			
• Total design codes	derived	25	Generalize structure in biology
Analogy (Design Codes)			
• Selected design codes	derived	140	Look for similar changes
Drill-Down (Design Codes)			
• Selected design codes	derived	183	Simplify complexity

Figure 1 indicates that the stereology literature database currently includes 60,000 data entries, of which more than half represent derived data. This community resource offers abundant opportunities for finding connections between and among the many parts that define biology.

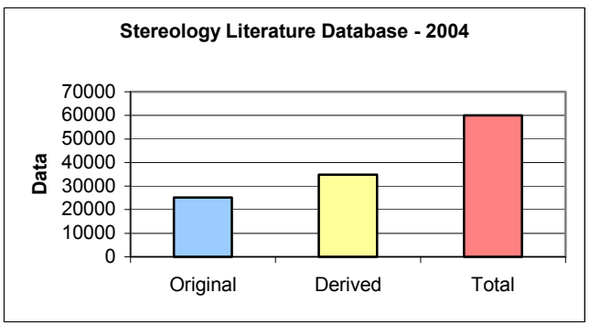


Figure 1. Research data stored in the stereology literature database.

Results: The principle findings of the project are listed below.

2001 to 2003

- Biological data can be transferred from research papers to a relational database and standardized.
- The production database demonstrates the feasibility of creating an electronic literature for the life sciences.
- A connection model for research biology yields widespread mathematical patterns, whereas the traditional change model does not.
- When stored in a database, published research data serve as a key resource for producing derived data.
- Biological data are subject to an ***uncertainty principle*** and therefore carry an unknown bias.
- Libraries can be designed that minimize bias (data pairs, design codes).
- Structures in biology are connected by rule (connection model).
- Algorithms can be written that generate organs and organisms from a single seed value.
- Complex research data can be unfolded by viewing data from a higher dimension.
- Relationships of structure to function can be expressed mathematically.
- Change in biology can be generalized and predicted.
- More than twenty thousand connections between structures in biology can be summarized by a single exponential equation.

New for 2004

- The organization of biological parts can be defined explicitly as repertoires of equations.
- The repertoire library views connections from different structural perspectives by forming networks of equations.
- Experimental data can be summarized by a single exponential equation, as shown earlier for control data.
- Mathematical analogies can encourage serendipity.
- Drilling down into a data set can reveal the presence of nested equations.
- Optimization may be a first principle of biology.

Repertoire Library (Data Pairs)

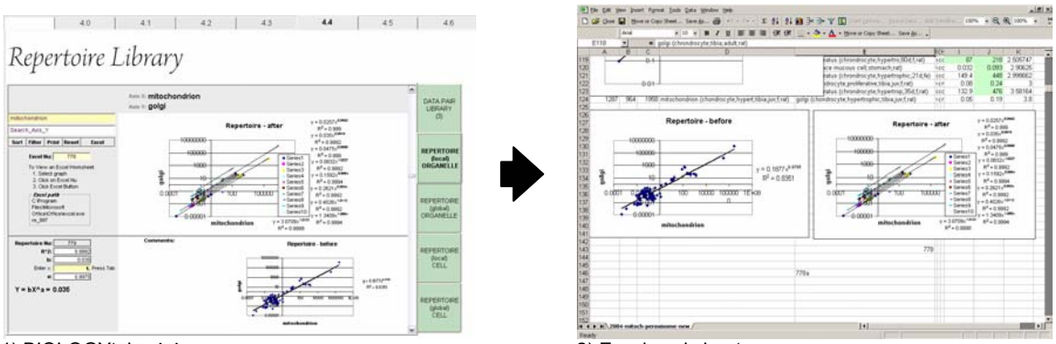
The repertoire library offers views of the same published data from many different perspectives by transforming tabular data into networks of equations. These networks, which show how structures are connected mathematically, are displayed as collections of power equations.

Organelles: If we imagine that each of the many parts of biology "sees" its world from a unique perspective, then what might we learn by sharing these different perspectives? For example, what parts of a cell would be of the greatest interest to a mitochondrion? One way of answering such a question is to list all those parts to which a mitochondrion has a mathematical connection. In other words, we can define a mitochondrial perspective as family of power equations ($y = bx^a$) wherein mitochondria (expressed as a volume or surface) would be the x variable:

$y(\text{organelle i}) = bx(\text{mitochondria})^a$.

Generating these equations is a simple task, using the literature database. Start with the data pairs table (included with the current EBS upgrade), type <mito> into the x name field, press Enter, click on the sort y radio button, and save the results as an Excel file. In Excel, notice that all the terms in the x name column begin with mito…, whereas the organelles in the y name column carry different names – sorted alphabetically. Start with Golgi and sort the x/y column numerically (low to high). Next, highlight the first three data pairs of Golgi and select a scatter graph. Change the x and y axes to logs and fit the points with a power regression line. If the r^2 is greater than 0.999, add extra points (row by row) to the graph until the r^2 comes close to 0.999. If not, move the calculation box down one row at a time until the r^2 becomes greater than or equal to 0.999. The power equation that appears on the graph describes a mathematical connection between

mitochondria and Golgi. Repeat the process for all the organelles and inclusions connected to mitochondria and the resulting set of equations becomes the mitochondrial perspective. (Note that for each graph, as shown in Figure 2, there is the corresponding Excel worksheet.)



1) BIOLOGYtabs 4.4. 2) Excel worksheet.

Figure 2. The repertoire library for mitochondria consists of a set of power equations. Calculations can be viewed by opening an Excel worksheet.

In turn, these equations – shown above as lines – can be programmed as computed fields on a work screen wherein a given value for a mitochondrion will generate the expected values for all the connected structures.

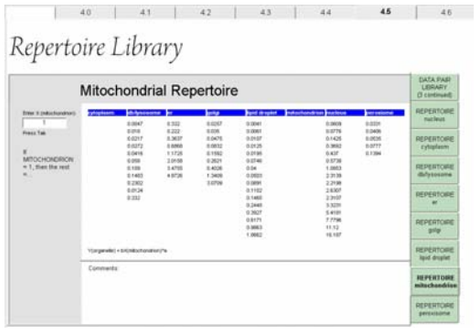


Figure 3. The repertoire library can also display the equations – illustrated as regression lines in Figure 2 – as an interactive network of equations. Apply a change to mitochondria and see how the connected organelles respond.

The first thing we can learn from this perspective is that mitochondria share connections with several cytoplasmic organelles. Notice, however, that mitochondria display multiple connections with the same organelle – each one being identified by a unique power equation. Similar patterns of order appear when perspectives are generated for the remaining organelles and inclusions. In other words, the proportion of organelles – one to another – typically appears as a collection of fixed relationships. In effect, the library includes a repertoire (catalogue) of mathematical connections for organelles that can occur across biology. It tells us that organelles are connected by rule and these rules can be captured by equations.

The repertoire library for organelles and inclusions contains 513 power equations typically with r^2s equal to or better than 0.999. If organelles maintain these stoichiometric-like proportions when an equilibrium state is reached after a change, then we can expect that a change in mitochondria will be accompanied by predictable changes in its connected neighbors. However, each of its neighbors – be it an organelle or inclusion – has its own set of connections. This means that an intricate web of connections could therefore translate the change we typically observe in a single organelle into a widespread effect. In such a setting, change would enjoy an unexpected wealth of complexity. Recall that complexity may be nature's way of producing new emergent properties (Bolender, 2001).

Notice that Figure 3 presents us with a challenging problem along with a new set of clues. If we imagine that the connections among the downstream products of a genetic

regulatory network can serve as quantitative markers of that network, then how would we write a set of equations describing the network for a specific cell in a specific setting? The solution to such a problem consists of selecting equations from each of the organelle columns that can describe – appropriately – a specific cell type in a specific setting. In turn, this connected set of equations would be expected to predict the relative proportions of organelles – in the specific cell. Our reward for solving this problem might be a compound mirror of equations with which to view earlier genetic activity – not unlike the way an astronomer looks at stars. The question, of course, is how do we select the appropriate equation(s) from each of the organelle columns?

The answer may be as simple as moving from one dimension to another. Consider the following experiment. Estimate the densities of the eight organelles in the specific cell type and calculate the ratio y/x for each data pair to get the proportions of the organelles (y(organelle) = bx(mitochondion)^{a}). Finally, compare the new proportions to those in Figure 3 and select the closest match. In effect, we can use Figure 3 as a lookup table. More importantly, perhaps, we now know how to turn a zero dimensional data point (y/x) – without connections – into a one dimensional power equation – with connections. In other words, by using the repertoire equations as a lookup table, we can transform an information poor data point into an information rich line – with surprisingly little effort. Indeed, such dimensional shifts may offer a host of new products and clues.

Cells: The repertoire library for cells identifies mathematical connections between cells and – as just described for organelles – the power equations display the step-like pattern. However, the cell-to-cell relationships are complex. The proportion of cells can be identified within and across species as (1) one cell to one cell or (2) one cell to many cells. For example, the proportion of pulmonary endothelial cells to type II cells in the mouse, goat and rat is the same (one to one), but this same proportion is also shared between endothelial cells and fat-storing, interstitial, Kupffer, macrophage, mesenchymal, and glial cells (one to many). The proportion is expressed as $y = 0.2693x^{0.9973}$, where x = endothelial cell and y = cell i. Such a result, which persists for many combinations of cells, suggests that proportions of cells are ordered by rule and that these rules are being conserved across species. Although the genetic mechanism responsible for controlling cell proportions is unknown, at least we now know that such proportions exist and that they can be quantified.

Global View: The repertoire library was also used to look for general patterns of organization. To generate global views for organelles and cells, the power equations were fitted to exponentials as described earlier for ladder equations (Bolender, 2003). When plotted as a group (without regression lines) they offer a striking view of organelle connections (Figure 4). Each blue point represents the y intercept of a power equation and each stack of points a different organelle view. The figure suggests that the cellular mechanism responsible for defining the organellar composition of cells operates by discrete steps (quanta) and that the underlying principles may be related, as suggested by the similar ranges and slopes of the exponential stacks. For those readers interested in exploring the organellar organization of cells, two questions quickly fall into focus. When and why do organelles locate at specific locations in the stack?

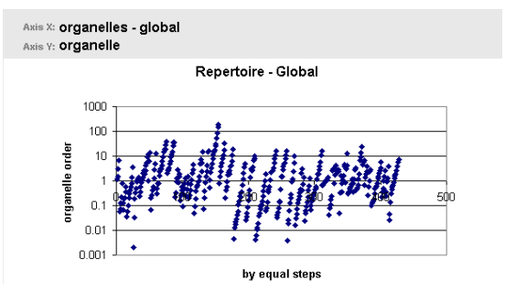


Figure 4. When the y intercepts of power equations are fitted to exponential equations, a global pattern of order can be seen. Connections between organelles appear to be ordered by rule.

Perspective: The repertoire library offers a global view of structural order by connecting local equations. It shows that different animals can produce remarkably similar parts, but

that the proportion of these parts – one to another – can be either similar or different. The fact that different animals can share similar connections and similar proportions of parts, suggests that they might also be sharing a similar genetic blueprint – or simply following a design strategy that leads to a similar phenotype. This means that each set of equations attached to a specific part reflects the general rules guiding the establishment of such relationships. In other words, the equations would seem to offer a broad overview of a fundamental organizing principle of biology.

If this proves to be the case, then it would not seem unreasonable to assume that we can assemble sets of equations for specific animals. Such an accomplishment would be accompanied by a substantial improvement in our ability to predict a large number of parts from one or a few seed values. Recall that the current form of the repertoire library makes no attempt to connect data across hierarchical levels – organelles and cells are treated separately. However, the only factor limiting such connections seems to be the amount of data available. In the future, large farms of equations spanning many hierarchical levels are likely to become commonplace.

Ladder Equation Library (Design Codes)

Last year, ladder equations were reported for the data pair library (control vs. control). By increasing the number of entries in the design code library (control vs. experimental) to 2,400, a similar – albeit provisional – estimate was also made for the experimental data.

If we start with the 2,400 design codes in the literature database, form ratios (structure y/structure x), sort the ratios (ascending), and collect sets of ratios that give power curves with an $r^2 = 0.9999$, we can generate a set of 23 equations describing the design codes. Since the slopes (a) of these power curves also tended to be close to one, the y intercept (b) of each equation served to identify a unit of order. In turn, when the y intercepts were plotted – as if they were rungs on a ladder – a ***single exponential equation*** of the form $y = e^{xa}$ – the ***ladder equation*** – appeared. This means that we can summarize the experimental data set of more than 2,400 entries with a single exponential equation having a $r^2 = 0.9991$:

$$y = 0.1194e^{0.1354x},$$

where y is the y intercept of the power equation and x the rung number.

Analogy Library (Design Codes)

In biology, interpreting the results of an experiment often includes reasoning by analogy. To wit, resemblances imply similarities. The analogy library takes this convention one step further by employing power equations as mathematical analogies. For example, if we detect a specific amount of change in a structure (e.g., mitochondria) and want to know where a similar change has occurred elsewhere, the library can provide such information.

Drill-Down Library (Design Codes)

Typically, a given design code is part of a larger code and, at the same time, consists of many smaller codes (Bolender, 2003). Recall that the design codes extend across the hierarchy of size as a set of nested equations. As such, they willingly serve as mathematical pathways to and from the genome.

The drill-down library illustrates that complex design code equations can be simplified by expressing them as two or more simpler equations (illustrated as before and after graphs). This coherency of equations is a fortunate relationship because it allows us to define an experimental process mathematically as the passage through a connected set of equations. In the drill-down library, the direction of information flow is from the organism to the gene –

across the hierarchical levels defined by the relational database model (Bolender, 2001b). Theoretically, a drill-down library can be used to find the genetic origin(s) of a biological part.

Methods and Results

Enterprise Biology Software (2004)

The Enterprise Biology Software package for 2004 updates the stereology literature database through 2003, adds the ***repertoire, analogy, drill down, and ladder equation (design code) libraries***, upgrades applications, and includes a progress report. Details of the upgrade can be found in the installation instructions (BIOLOGYtabs 2004).

Stereology Literature Database

Database Update: This year, data taken from submitted reprints were added to the literature database and design codes were harvested from an additional 165 papers.

Libraries

Previous libraries were updated to reflect the recently entered data and new libraries were generated from the data pair and design code files.

Searching Libraries: The data pair library includes two columns of control data, whereas the design code library includes one column each of control and experimental data. The major purpose of these libraries is to generate equation libraries. Recall that:

Data Pairs (control vs. control)

- Detect a connection between two structures, two functions, or a structure and a function.
- Detect connections among several structures, functions, and structures and functions.
- Compare control data coming from one or several papers.
- Generate equations for predicting structure and function.
- Identify patterns.
- Define repertoires for organelles and cells with equations.

Design Codes (control vs. experimental)

- Detect change ***quantitatively*** and ***qualitatively*** as connected sets.
- Identify patterns of change.
- Generate equations for predicting changes in structure and function.
- Use equations to search for analogies.
- Offer a drill-down approach for identifying nested equations.

To simplify their use, both data pair and design code libraries share a similar interfaces and methods for generating equations (Figure 5).

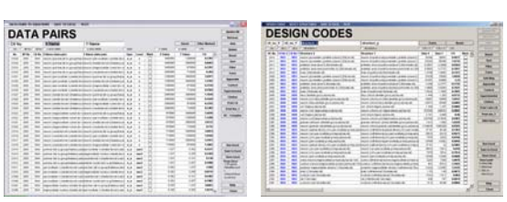


Figure 5. Data of the stereology database are selected from tables and sent to Excel

worksheets where they can be fitted to regression equations.

Data Pair Library

Types: The ***data pair library*** (BIOLOGYtabs 2004; 4.2-4.6) includes sets of equations calculated as regression curves – derived from the data pair table (Figure 5). An equation defines a quantitative connection between two or more parts. This library includes collections of data and equations.

1. ***Data Pair Library – Find Equations***: The data table (shown in Figure 5) is used to generate data pair equations.
2. ***Data Pair Library (Connection Equations)***: Identifies quantitative relationships of structure to structure, structure to function, and function to function.
3. ***Data Pair Library (Ladder Equations)***: Summarizes control data pairs locally (power equations) and globally (exponential) equations.
4. ***Data Pair Library (Repertoire Equations)***: Displays connections calculated from the unique perspectives of named organelles and cells.

Properties and Rules: See Progress Report 2003 (Bolender, 2003).

Applications: Tab 4 of BIOLOGYtabs 2004 displays the data pairs as a table of X and Y values (4.1), a collection of related structures (4.1), a collection of ladder equations (4.2), a repertoire of connections for organelles and cells (4.3), and networks of equations link (4.4, 4.5)

- ***4.1: Data Pair Library – Find Equations:*** Click on the picture button to display the data pairs table (Table 6). Use the table to select structures of interest and then export them to an Excel file for analysis.

Figure 6. Finding equations.

- ***4.2: Data Pair Library:*** The collection illustrates – with equations – pairs of structures connected by rule (Table 7). Notice that mathematical order can be maintained within and across both structures and species – at one or more levels. In effect, the data pair library helps to explain the organizational complexity of biology in terms of equations. Excel worksheets, which show how the equations were generated, are readily available for viewing.

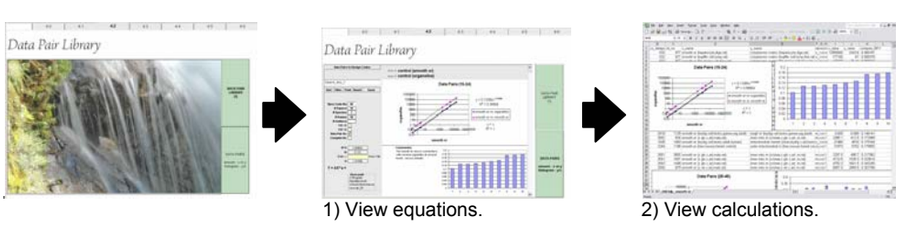


Figure 7. Viewing data pair equations.

- ***4.3: Ladder Equation Library:*** The library shows that biological data can be ordered locally (power equations) and globally (exponential equations) and that the order occurs in distinct steps (quanta). The steps can be seen as parallel sets of equations and as steps in histograms (Figure 8). The library includes ladder equations for the total collection of data pairs and for subsets thereof (organ, cell, and organelle).

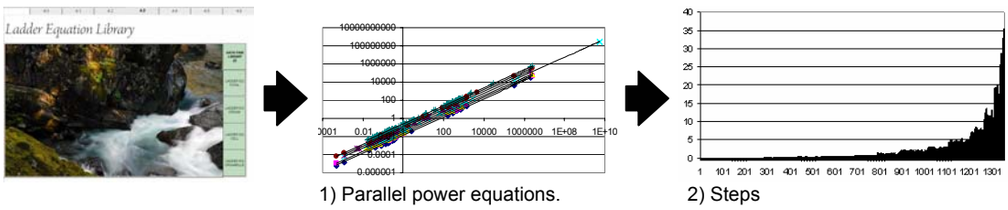


Figure 8. Rung equations appear as sets of parallel curves and as steps in histograms.

4.3: Repertoire Library Equations: The library illustrates the types of connections being used by organelles and cells. In effect, it can show – with equations – the what, where, or when an animal genome makes structures. It also illustrates – with before and after graphs – the process of extracting order (as equations) from otherwise noisy data (Figure 9).

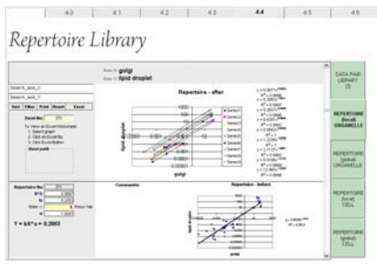


Figure 9. Repertoire equations.

Repertoire equations are displayed in both local and global settings. Local equations offer a detailed picture of connections between specific data pairs, whereas the global equations summarize many local equations with a single exponential equation.

4.5, 4.6: Repertoire Library Equations (expressed as a network): If we consider each local repertoire equation to be a piece of a biological puzzle, then connecting these pieces mathematically gives us a networked view. Expressing the equations of 4.3 as columns of computed fields provides a repertoire of networked equations. In the nuclear repertoire (Figure 10; right), notice the white data entry field at the left. When a value for the nucleus is entered, all the numbers in the organelle columns at the right change by rule (as defined by the repertoire equations).

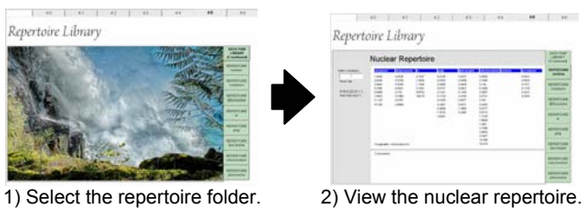


Figure 10. Repertoire equations as networks.

Notice that for each nuclear value, there are 8 values for cytoplasm, 8 for dense bodies, 6 for endoplasmic reticulum, 11 for Golgi, 10 for lipid droplets, 18 for mitochondria, and 6 for peroxisomes. In other words, change the volume of the nucleus and all the cytoplasmic structures with a mathematical connection to the nucleus change their volumes by rule. This tells us that our ability to design a prediction model for a specific cell type in a particular setting depends – importantly – on knowing how to pick the appropriate equation(s) from each organelle column.

Design Code Library

Types: The ***design code library*** (BIOLOGYtabs 2004; 5.2-5.6) compares control to experimental data with power equations. The library includes six collections.

1. ***Design Code Library – Find Equations***: A data table is used to generate design code equations.

2. ***Design Code Library (Simple Design Codes)***: Identifies patterns of change – one paper at a time.
3. ***Design Code Library (Complex Design Codes)***: Identifies patterns of change – several papers at a time.
4. ***Design Code Library (Ladder Equations)***: Summarizes control and experimental data locally with power equations and globally with exponential equations.
5. ***Design Code Library (Analogy Equations)***: Identifies similar changes as those data that share the same regression equation.
6. ***Design Code Library (Drill-Down Equations)***: Identifies patterns in complex data and provides a strategy for tracking changes back to the genome.

Properties and Rules: See Progress Report 2003 (Bolender, 2003).

Applications: Tab 5 of BIOLOGYtabs 2004 displays the design codes as a table of x and y values (5.1), a collection of simple code equations (5.2), a collection of complex codes (5.3), ladder equations (5.4), analogy equations (5.5) and drill-down equations (5.6).

- ***5.1: Design Code Library – Find Equations:*** Click on the picture button to display the design code table (Figure 11). Use it to select structures of interest and then export them to an Excel file for analysis. Data previously used to generate an equation can be viewed by opening its Excel worksheet.

Figure 11. Design code library – Find equations. Select data > send to worksheet > make calculations.

- ***5.2: Design Code Library (Simple):*** The library of equations is grouped according to control, experimental, development, aging, disease, and relationships of structure to function (Figure 12). The design code equations were calculated using data from a single research paper.

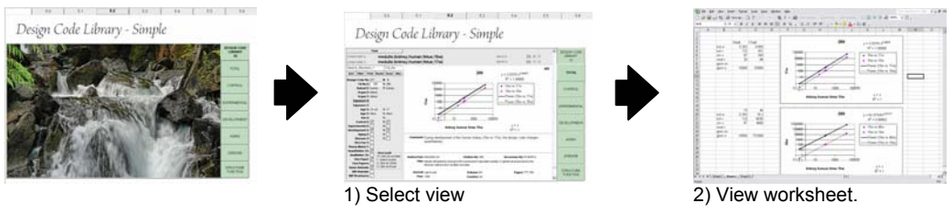


Figure 12. Design code library – Simple.

- ***5.3: Design Code Library (Complex Equations):*** A complex design code combines the data of several simple design code equations (Figure 13). It characterizes change by structure and by event. Recall that structural data (V, S, L, N) can be used to identify both ***qualitative*** and ***quantitative*** changes, whereas density (Vv, Sv, Lv, Nv) and mean (mV, mS, mL) data can detect ***qualitative*** changes.

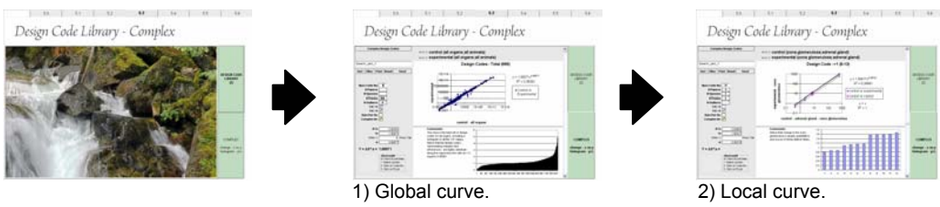


Figure 13. Design code library – Complex.

Types: The ***ladder equation library*** (BIOLOGYtabs 5.4) includes a collection of ladder (exponential) and rung (power) equations that together summarize data for total and selected sets. The summary takes the form of a single exponential equation: $y = e^{xa}$.

Properties: The ladder equation summarizes all the experimental connections in the library database (Figure 14) – expressed as design codes (Figure 12) – with the single expression:

$y = 0.1194e^{0.1354x}$,

where y equals the y intercept of the power (rung) equations and x the number of the rung (e.g., 1 to 24).

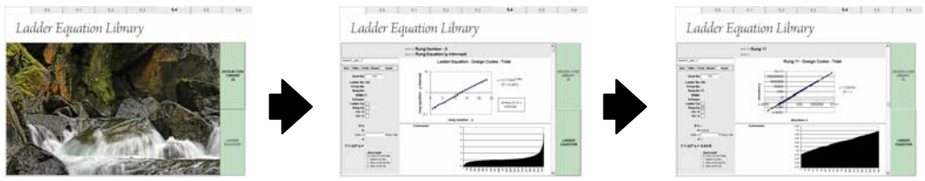


Figure 14. Ladder equation library for design codes.

For a discussion of ladder equations, see Bolender (2003).

5.5: Design Code Library (Analogy Equations): The library includes a collection of equations that plot experimental data for the same structures coming from different experimental settings and animals (Figure 15).

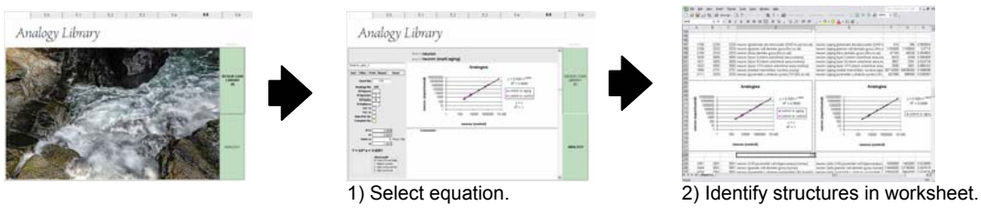


Figure 15. Analogy equations find structures that change similarly.

Types: The ***analogy library*** (BIOLOGYtabs 5.5) was derived from the design code library and includes one collection.

1. ***Design Code Library (Analogy)***: Identifies structures that display similar changes – under similar and different conditions.

Properties: Analogy equations display several properties.

1. When the sets of curves are parallel and separated, they show an increase or decrease in the absolute amount of same material – a **quantitative** change. Most items fall into this category.
2. In contrast, nonparallel curves suggest a change in the proportion of the parts – a **qualitative** change. Details of the calculations can be found in Excel worksheets.

Application: The library continues to reinforce the view that the composition of cells and the changes therein are tightly controlled by the genome and that this control often extends across the species boundary.

5.6: Design Code Library (Drill-Down Equations): The library includes a collection of equations that identify patterns in complex data and contribute to a strategy for tracking changes back to the genome (Figure 16).

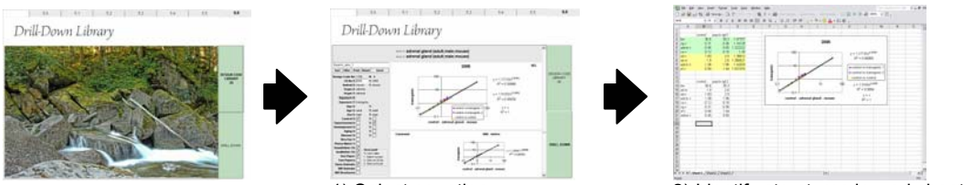


Figure 16. The drill-down library finds order in research data.

Types: The ***drill-down library*** (BIOLOGYtabs 5.6) was derived from the design code library and includes one collection.

1. ***Design Code Library (Drill-Down)***: Identifies underlying order in complex research data taken from a single paper.

Properties: The drill-down library illustrates the properties of nested equations.

1. A complex data set can be simplified by fitting the data to several power curves.
2. Unfolding complexity improves the resolution of data analysis.
3. Drilling-down into equations can occur within and across all levels of the biological hierarchy of size.

Application: The library uses a before and after format to illustrate how a complex data set can be simplified by resolving it into two or more power equations, each with a $r^2 \geq 0.999$. If these equations reflect the underlying order of a cellular response, then they suggest that the changes we typically report may in fact be a composite of several nested changes. To wit, order in biology can be defined quantitatively as equations embedded in equations.

Discussion

Turning Papers into Equations

The **Enterprise Biology Software Project** turns reprints into equations and equations into discovery platforms. Such an outcome owes its success to the stereology literature, which is an ideal place to hunt for order in biology. Although such hunting requires assistance from mathematics and technology, these research tools can be readily captured in software and distributed freely to contributing authors.

The equations tell us that mathematical order in biology can be found in the connections between structures and that the connections define networks. This universal connectivity creates robust pathways for both discovery and prediction. Moreover, the connections support the dimensional shifts that provide access to otherwise inaccessible information.

Equations ease the discovery process – enormously – because they can simplify complexity. Consider, for example, the repertoire equations in BIOLOGYtabs (4.4-4.6). They demonstrate that the relationship of one structure to another is complex, but that the complexity can be explained by a set of equations. Once a pattern of connections is found, hunting for the next level of complexity consists merely of generating yet another equation library. One result leads naturally to the next. What could be simpler?

Biological Stereology and the Systems Biology Revolution

The **Systems Biology Revolution** promises to translate the results of the genome sequencing projects into an understanding of how genes produce and maintain the all the many structures and functions that are biology. This includes the daunting task of unraveling the complexities of the genetic regulatory networks. Since the reach of these networks extends across the biological hierarchy of size, they can be studied by starting at either end or somewhere in between – with the option of proceeding upstream or down. Success in mapping and explaining these networks will no doubt require massive amounts of highly reliable data supported by robust connections.

In figuring out these regulatory networks, we can imagine two likely directions for research – upstream and down. One seems easy the other hard. When moving upstream from the gene products to the genes, all the organizational information is already in place and just waiting to be discovered. Moreover, moving upstream is equivalent to going from a complex setting to a simpler one, a task easily handled by equations (e.g., the repertoire and drill-down libraries). Moving downstream, however, is more challenging because each item added to the network must be fitted painstakingly into the scheme qualitatively and quantitatively – one step at a time – with the appropriate authentication. In short, it is often easier to take something apart than to put something together – especially in the absence of a blueprint.

Biological stereology can contribute importantly to systems biology. It generates critical data sets routinely, moves data throughout the structural hierarchy with ease, identifies specific structures at specific locations in intact animals with light and electron microscopy, and regularly leverages its relational database of published research.

Is there a ready link between stereology and molecular biology? Yes, indeed. Since the data of molecular biology can be expressed as ratios, we have a direct match with the data pairs and design codes described herein. How might this be helpful? If, for example, we combine the ratio data of the stereology database with those of microarray analyses in an Excel worksheet, then specific genes – or gene products – could be run with the stereological data to hunt for regression curves with $r^2 \geq 0.999$. Finding such curves might point to associations between downstream structures and their upstream genes or gene products. Such an approach may prove especially helpful because the products, locations, and functions of so many genes remain a mystery. Since the results of many microarray experiments are already being published in catalogues on the

Internet, large amounts of gene expression data are freely available to stereologists. An introduction to these new microarray methods may be of interest to the reader (www.ncbi.nlm.nih.gov/About/primer/microarrays.html .

As the systems biology revolution takes root, we will follow with great interest the strategies developed for dealing with the requirements of a critical data set. On close inspection, it appears that many papers in molecular biology rely exclusively on density data for detecting differences and changes. Such a narrow approach may hinder even the most determined efforts at unraveling complexity. Piggybacking on the strengths of stereology may therefore offer a welcome and profitable solution. Such a strategy makes sense. Stereology offers a mathematical infrastructure capable of moving data across the biological hierarchy of size. Combine microarray data with those of stereology and the mathematical connections extend all the way from genes to organisms – and back.

Unfolding Complexity with the Equation Libraries

Since the products of genetic regulatory networks include the downstream expression of molecules, organelles, and cells, working our way upstream from these structures should take us back to the origins of these networks. In effect, quantifying the end products of the genetic regulatory networks at different hierarchical levels of expression becomes an exercise in unfolding complexity. This unfolding process might also tell us something about the rules that are in play.

Within the framework of a connection model, complexity in biology can be unfolded using two simple rules. Consider the structures, A and B (Figure 17). They are connected by rule – the **proportion rule**. However, the proportion rule can vary according the amount of each structure A and B, according to the **absolute amount rule**.

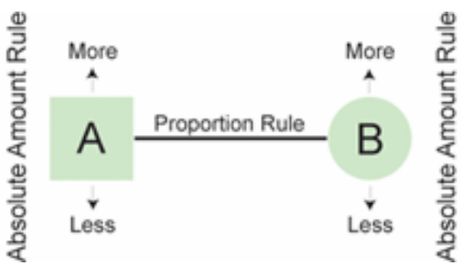


Figure 17. Rules for detecting differences and changes.

The ratio of the two structures (A/B) defines the connection of a data pair or a design code. More or less of structures A and or B will change the proportion rule, except when both A and B change equally. In this case, A/B (green) = A/B (blue) = A/B (red), as shown in Figure 18. Recall that data related to a structure, to an average structure, and to a unit of reference volume (a density) can be used to calculate the ratio of two structures. However, when only density data are available for this calculation, only part of the story will be known. The contribution of the absolute amount rule will remain a mystery (see Table 2).

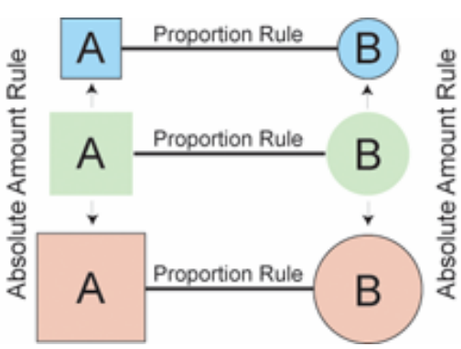


Figure 18. Rules for detecting differences and changes. In this example, the

proportions are the same, even when the absolute amounts are different.

This means that density data alone cannot detect the absolute amounts of the structures. Since genetic regulatory networks control both the proportions of structures and their absolute amounts, the complexity of a connection between two structures can be reduced to two rule-based events. The **proportion rule** can be detected with just density data, but the **absolute amount rule** requires the information of a critical data set. Both are equally important for explaining a connection. Before leaving this figure, there is an additional point worth mentioning. Recall that the majority of the differences and changes detected with the design code equations show experimental curves parallel to the reference curves (BIOLOGYtabs 5.0). This means that the experimental structures fitted to the regression line reacted uniformly as a group, as illustrated by the red and blue structures in Figure 18. In such a case, the **proportion rule** would indicate no difference or change, whereas the **absolute amount rule** would tell a very different story. In other words, density data alone cannot detect an increase or decrease when the curves are parallel.

If we start with two structures A and B and introduce a difference (more or less), then we can see that many outcomes exist.

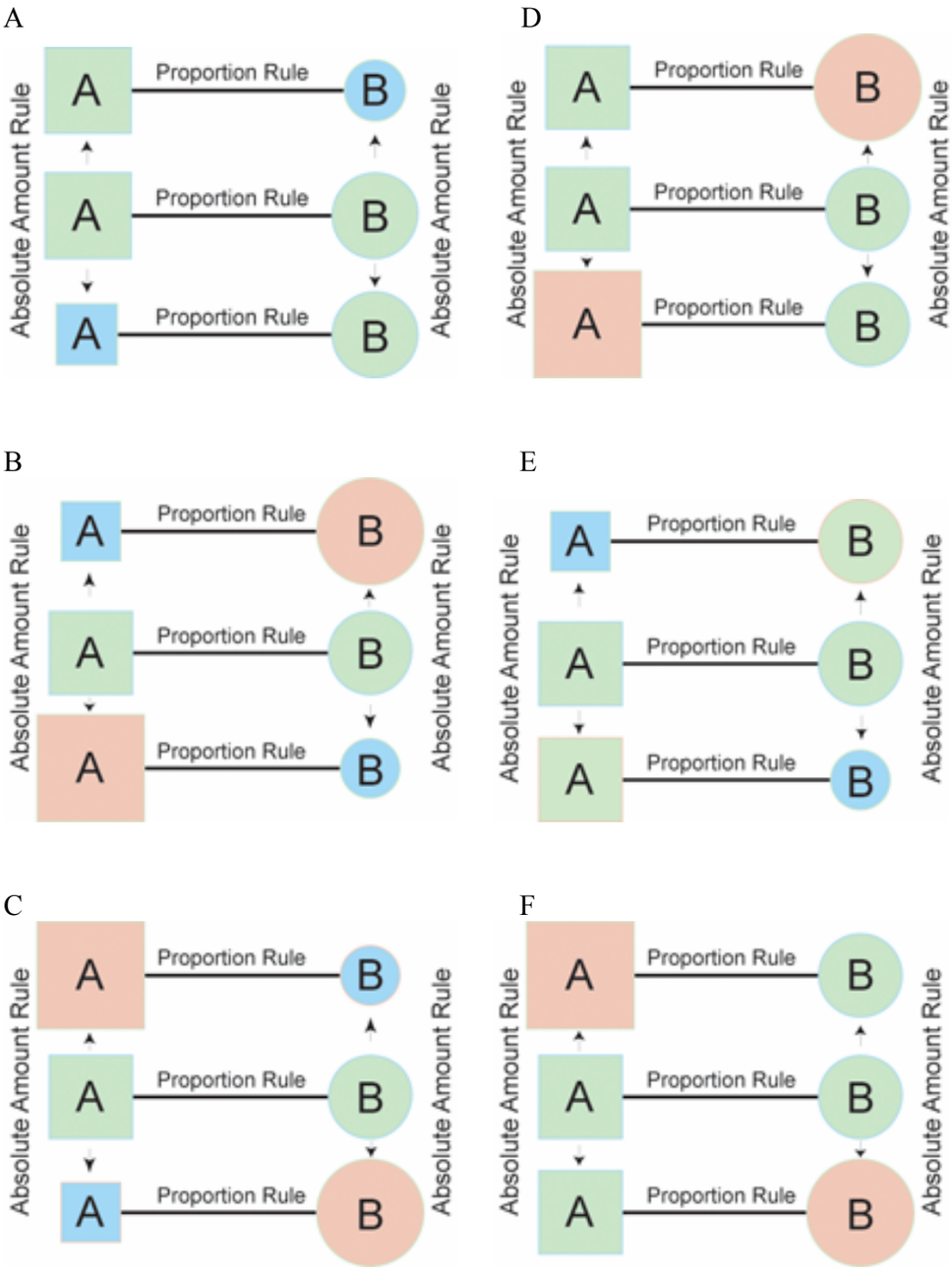


Figure 19. Detecting differences and changes is a function of two rules.

In turn, we can summarize Figures 18 and 19 with Table 2 by expressing the connections as magnitudes (absolute amounts) and directions (proportions =1, >1, <1). Note that B = blue (less), G = green (same), and R = red (more). The table demonstrates that the nine absolute events produce three relative outcomes.

Table2. Detecting differences and changes using a rule-based system (see Figures 18, 19).

Figure	Pattern	Absolute Amounts A/B	Interpretation (Abs Amt Rule) A B	Connection A/B	Proportion	Interpretation (Proportion Rule) A/B
18	**BBGGRR**					
	BB	50/50	↓	1.0	=1	=
	GG	70/70	=	1.0	=1	=
	RR	90/90	↑	1.0	=1	=
19A	**GBGGBG**					
	GB	70/50	= + ↓	1.4	>1	↑
	GG	70/70	=	1.0	=1	=
	BG	50/70	↓ + =	0.7143	<1	↓
	GRGGRG					
	GR	70/90	= + ↑	0.7777	<1	↓
	GG	70/70	=	1.0	=1	=
	RG	90/70	↑ + =	1.2857	>1	↑
	BRGGRB					
	BR	50/90	↓ + ↑	0.5555	<1	↓
	GG	70/70		1.0	=1	=
	RB	90/50	↑ + ↓	1.8	>1	↑
	BGGGGB					
	BG	50/70	↓ + =	0.7143	<1	↓
	GG	70/70	=	1.0	=1	=
	GB	70/50	= + ↓	1.4	>1	↑
	RBGGBR					
Control	RB	90/50	↑ + ↓	1.8	>1	↑
	GG	70/70	=	1.0	=1	=
	BR	50/90	↓ + ↑	0.5555	<1	↓
	RGGGGR					
	RG	90/70	↑ + =	1.2857	>1	↑
	GG	70/70	=	1.0	=1	=
	GR	70/90	= + ↑	0.7777	<1	↓

Taken together, the figures (18, 19) and the corresponding data table (Table 2) summarize the complexity of a connection between two structures - A and B. This same pattern applies to each pair of structures at every level of the biological hierarchy of size. Observe, however, that the complex relationship of structure A to structure B can only hint at the real-world complexity wherein many more than two structures are connected (Figure 3).

The point of the example is to show that a difference or change in a connection can be produced by more than one event. For example, no change in structure A and a decrease in structure B produce a **relative increase** (Figure 19 A). Moreover, notice in Table 2 that a **relative increase** can be explained by three very different "genetic" events. This means that densities – by themselves – cannot detect a gene based event reliably because such an event involves the expression of two rules – not one. Examples of these rules actually being expressed can be found throughout the derived data libraries.

Progress as a Progression of Models

One way of assessing the progress of the **Enterprise Biology Software Project** is to review how the stereology literature database can be analyzed by applying different data models. Notice how each model attempts to deal with complexity and how the solutions help to define the subsequent model.

Access Model: Information stored at many different locations (journals) was transferred to a single location – a relational database. In turn, the database and supporting software were distributed freely to the user community.

- Information can be found in many different ways by searching on the columns of data tables, as defined by the database model.
- Data can be reformatted (standardized papers) and used to generate new derived data libraries.
- A single database model can be designed to accommodate most types of biological data.

Change Model: Research data can be can be expressed as a percentage change by dividing an experimental value by its control and then multiplying by 100%.

- In biology, a critical data set is required to detect a change unambiguously. It includes the volume of a structure, the number of cells in the structure, and the density of the part being followed.
- The change model can be used to demonstrate the validity of the critical data set requirement by illustrating that density data alone cannot detect change unambiguously. This point can be confirmed quickly by comparing density data to structure data, using the color-coded screen (BIOLOGYtabs 3.4).
- Although the standardized data can be connected hierarchically, the data of a change model cannot be converted readily into libraries of local and global equations. See Appendix > Literature Database > View New Data > Phenotype Data.

Connection Model: The connection model relies on the quantitative relationship of two structures, which includes data pairs (for control data) and design codes (for experimental data).

- Connection data can be converted routinely into libraries of local and global equations.
- By forming data ratios, the connection model minimizes experimental bias.
- Complex data can be unfolded and then refolded.
- Change can be separated into qualitative and quantitative components.
- Control and experimental data sets can be summarized – individually - with a single exponential equation.
- An analysis of connection data reveals distinct patterns of order in biology

Network Model: The network model connects the equations of the connection model into local and global networks.

- Algorithms can be written for organs and organisms. They demonstrate that data can be generated – upstream and down – from a single seed value.
- A repertoire network (BIOLOGYtabs 4.4-4.6) defines the mathematical relationships within a collection of connected structures. In turn, local networks can be connected to form global ones (work in progress).

A First Principles Approach

A first principle can be defined as a law upon which others are founded or from which others are derived. It is a general truth, comprehending many subordinate truths, and not deductible from others.

How do we hunt for first principles in biology? Consider the strategy of the ladder equations wherein all the connections between structures were generalized with a single equation – one for control data and another for experimental. However, finding two equations instead of one suggested that the equations must be subordinate to a larger truth. Notice what happens when we plot the two equations together (Figure 20). The curves intersect and bear a striking resemblance to the solution of a linear programming problem. Is this telling us that the structure of living things is built on a mathematical platform designed to produce the best results? If the answer is yes, then we may have a candidate for a first principle (Walthrop, 1992).

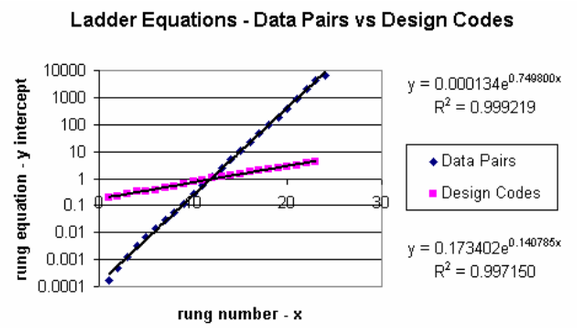


Figure 20. Is structural organization a best solution?

Figure 20 presents us with a new challenge. Starting with these two equations, can we reverse the summarizing process and tease out the details of structural order - across animals and experimental settings? In other words, can we write networks of repertoire equations capable of predicting the structure of specific animals in specific settings?

Such an exercise may one day become a practical necessity. We now know from genome sequences that organisms can have remarkably similar genotypes, but amazingly different phenotypes. If we begin with such a similar genetic blueprint, why do we end up so different? If, for example, we share 97% of our genes with the chimpanzee, does this mean that only 3% of our genome is uniquely human? If we don't have species-specific genes, what do we have? If, instead, we discover that speciation depends importantly on the types of connections that form between structures, should we be looking for "structural compounds" with unique properties? If such compounds exist, can we expect to find stoichiometric-like rules operating at and across all levels of organization?

Concluding Comments

We now have a better understanding of why published data in their original form fail to yield widespread patterns of mathematical order in biology. Biological data carry an unknown bias and the same structural part can assume many different values in different settings. A principal observation of the ***Enterprise Biology Software Project*** is that the mathematical organization of biology occurs in the connections between the structural parts and that these connections can be captured as libraries of equations. Indeed, the equations reveal a pervasive mathematical order both within and across species. By embracing structures of all sizes from all animals, the relational database model – based on the principles of stereology - can provide the local and global views of biology fundamental to the task of unraveling complexity. Although the magnitude of biological complexity remains undetermined, we now know how to unfold complexity and to explore it with equations. In short, we are learning how to explore biology as a mathematical puzzle.

References

Bolender, R. P. 2001a Enterprise Biology Software I. Research (2001) In: Enterprise Biology Software , Version 1.0 © 2001 Robert P. Bolender

Bolender, R. P. 2001b Enterprise Biology Software II. Education (2001) In: Enterprise Biology Software , Version 1.0 © 2001 Robert P. Bolender

Bolender, R. P. 2002 Enterprise Biology Software III. Research (2002) In: Enterprise Biology Software , Version 2.0 © 2002 Robert P. Bolender

Bolender, R. P. 2003 Enterprise Biology Software IV. Research (2003) In: Enterprise Biology Software , Version 3.0 © 2002 Robert P. Bolender

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