Enterprise Biology Software: IV. Research (2003)

ROBERT P. BOLENDER

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

Summary

Among the many challenges facing research biology today, we can identify at least three that promise to accelerate discovery in the near future. These include linking diverse information across a biological hierarchy of size, identifying underlying principles of biology, and transforming biology into an information science. Using the stereology literature as a well-spring, the ***Enterprise Biology Software Project*** actively encourages investigators pursuing such challenges by making new information and technologies freely available. The process is simple and direct. First authors of papers applying stereology, as listed each year in ***PubMed*** (National Library of Medicine), are invited to submit reprints as candidate papers for the stereology literature database. Once entered into the database, these research data can be accessed as is or used to generate libraries, equations, patterns, platforms, or whatever one might wish. At the beginning of the year, the updated databases, libraries, and software tools are distributed to contributing authors – past and present - on a CD. This release includes the updated stereology literature database (research data through 2002), new libraries (***design code***, ***ladder equation***), a progress report, and several unexpected findings.

Introduction

Background

What have we learned so far? We now know it is possible to standardize most types of biological data with a relational database, using a mathematical organization based on stereology. In effect, by creating a production database for the literature of biological stereology, we have demonstrated a core facility for research biology and tested the feasibility of producing and distributing an electronic literature (Bolender, 2001a, 2001b, 2002). However, a central challenge of the database exercise was to address – aggressively – the very real problem of complexity in biology. To begin, the problem of understanding complexity was divided into three tasks: (1) organizing published research data by combining three models (qualitative, quantitative, and relational), (2) unfolding complexity into elements by identifying distinct sources (biology, methods), and (3) recombining the elements to look for patterns and underlying principles.

Using this approach, we have seen that our experimental methods of extracting data from biology introduce an ***uncertainty principle*** in that all or most of our stereological estimates carry an unknown bias. In turn, we considered the practical implications of bias and showed how to minimize its effects (Bolender, 2002). The exercise using biological algorithms to predict data to and from the genome – from a single seed value – made the indelible point that a believable diagnosis and prediction system would have to be based on equations displaying coefficients of variation equal to one (R^2=1). This observation was particularly important because it provided the incentive for assembling two new libraries that offer opportunities well beyond those of diagnosis and prediction. For example, we can now look at change from a higher dimension and begin to understand key elements of its complexity. This recent progress with structural data establishes guidelines for the far more difficult task of connecting the data of stereology with those of biochemistry and molecular biology.

Progress

The ***Enterprise Biology Software Project*** follows a data driven route to discovery. Data are taken from research articles, stored in a relational database, and standardized. In turn, these original data are used to generate derived data. The two main products of the stereology literature database thus far include (1) a collection of data libraries and (2) the findings generated by these libraries. The discovery strategy is simple: ***find order and follow it***.

Libraries: Libraries serve as discovery platforms (Table 1). They include one or more user interface screens, data, help files, and often worked examples (e.g., Excel scratch sheets; case studies).

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	14,290	
• Experimental Data	original	9,386	
• Contl data – by data point	original	11,051	Find data by data point; level
• Contl+Exptl data – by data point	original	6,438	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	21,035	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
Design Code			
• Local (data from 1 paper)	derived	880	Identify and predict change
• Global (data from1+ papers)	derived	58	Identify and predict change
Ladder Equation			
• 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

Figure 1 indicates that the stereology literature database currently includes 55,000 data entries, of which more than half represent derived data. This resource offers abundant opportunities for finding connections between and among the many parts that define biology.

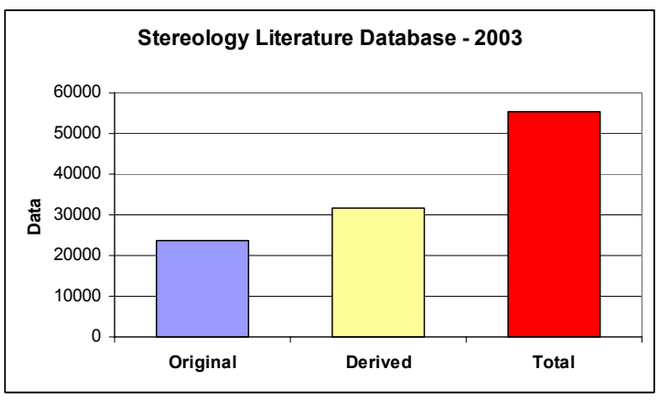

Figure 1. Data in the stereology literature database.

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

- 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.
- When stored in a database, published research data serve as a key resource for producing derived data.
- Since biological data are subject to an ***uncertainty principle,*** they carry an unknown experimental bias.
- Libraries can be designed that minimize bias (data pairs, design codes).
- Structures in biology are connected by rule (connection model).
- Algorithms can generate organs and organisms from a single seed value.
- Sources of complexity in research data can be identified by viewing data from a higher dimension.
- Relationships of structure to function can be expressed mathematically.
- Change in biology can be generalized and predicted.
- Twenty thousand connections between structures in biology can be summarized by a single exponential equation.

Design Codes

To apply information technologies to biology, we need to understand how biology manages information. We know that DNA stores information in genes that can be translated with the help of RNA into protein molecules, etc. If we imagine that this process of distributing information continues in an orderly way, gaining richness and complexity in forming all the many parts of a living system, then it seems likely that all the steps of the process are connected. In other words, the information and its expression across the biological hierarchy must be nested hierarchically - according to well defined design principles.

To simplify the task of identifying and quantifying connections in biology, we can design a new library consisting of ***design codes***. A design code can be defined as an equation – or a set of equations - that represent rules for connecting the parts of a structure. Moreover, we can assume – for convenience - that design codes are nested hierarchically everywhere – from molecules to organisms. From this definition it follows that a given design code is part of a larger code, while at the same time it contains many embedded codes.

How can we use design codes? They allow us to observe – in greater detail - the behavior of change in living systems. In addition to supplying local information (qualitative; quantitative), design codes also identify global patterns of change that appear when several codes are combined across publications and animals (complex codes). The design code library even supplies a paradoxical view of change – one in which change becomes a "***constant.***" In effect, design codes suggest that change operates by rule and behaves in a predictable way.

Last year (Bolender, 2002), we observed that sets of equations capable of predicting structure and function from a single seed variable required a set of equations with an $R^2=1$. In the real world, of course, such a requirement becomes impractical. If, however, we relax the requirement only very slightly to $R^2 \geq 0.999$, then we can extract many design codes from the stereology literature. Here the strategy consisted of removing outliers from a data set until the $R^2 \geq 0.999$. In this case, an outlier was defined as a point that did not fall on or next to a regression curve. Details of this harvesting process can be seen in the Excel scratch files that are included with the software upgrade.

Ladder Equations

When exploring biology as an information science, one strategy to follow consists of finding order and then tracking the order to its source. Ladder equations serve as another

example of this process. If we start with the 20,000 data pairs 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=0.999, we can generate a set of 24 equations describing the 20,000 data pairs. Since the slopes (a) of these power curves tend to be close to one, the y intercept (b) of each equation can serve to identify a unit of order. In turn, when the y intercepts are plotted - as if they were rungs on a ladder – we get a ***single exponential equation*** of the form $y=e^{xa}$ – the ***ladder equation***. Repeat this process, but restrict it to organs, cells, or organelles and we discover additional ladder equations. In effect, order in biology appears as equations embedded in equations. This observation will be of interest to us shortly when we turn our attention to making speculations about how genes might be operating.

Methods and Results

Enterprise Biology Software (2003)

The Enterprise Biology Software package for 2003 updates the stereology literature database through 2002, adds the ***design code and ladder equation libraries***, upgrades applications, and continues to explore complexity.

Stereology Literature Database

Database Update: This year, data from about 400 publications were added to the literature database.

Relaxing the Data Entry Rule: Since the stereology literature database is based on a change model, only those papers meeting the requirements of a ***critical data set*** were selected for data entry (Bolender, 2001a). Derived data, however, allow us to change the rules. By upgrading the literature database from a change to a connection model, papers reporting only density or mean cell data can now be added to the database and used ***selectively*** to hunt for patterns. This means that a much larger proportion of the stereology literature can now be used to generate derived data.

Relaxing the R^2=1 Requirement: In an ideal world, the equations of prediction models would have R^2s equal to one. If - in our world - we relax this requirement by only one tenth of one percent, then the stereology literature yields many design code equations. The question, of course, remains the same. How good are the equations at predicting change? If 100% represents a perfect outcome, then the observed mean score of 100.4% missed the goal by less than one half of one percent (N.B., the standard deviation was 6.6 and the number of samples 880). The results would be slightly better if data entry had been strictly limited to equations with $R^2 \geq 0.999$. Several entries with R^2s of only 0.99 were allowed for purposes of illustration. In any case, the results suggest that change can be predicted – remarkably well – with the design code equations.

Libraries

All the previous libraries were updated to include the newly entered data and two new libraries were added (design code and ladder).

New Strategy for Searching Libraries: The data pair and design code libraries offer ready access to equations with $R^2 \geq 0.999$. Recall that:

Data Pairs (control vs. control)

- Detect a connection between two structures, two functions, or a structure and a function – by one or more papers.

- Detect connections among structures, functions, and structures and functions – by one or more papers.
- Compare control data across several papers.
- Generate equations for predicting structure and function.

Design Codes (control vs. experimental)

- Detect change ***quantitatively*** and ***qualitatively*** as connected sets - by one or more papers.
- Identify patterns of change.
- Generate equations for predicting change in structure and function.

To simplify their use, both data pair and design code libraries share a similar interface and method for generating equations with $R^2 \geq 0.999$ (Figure 2).

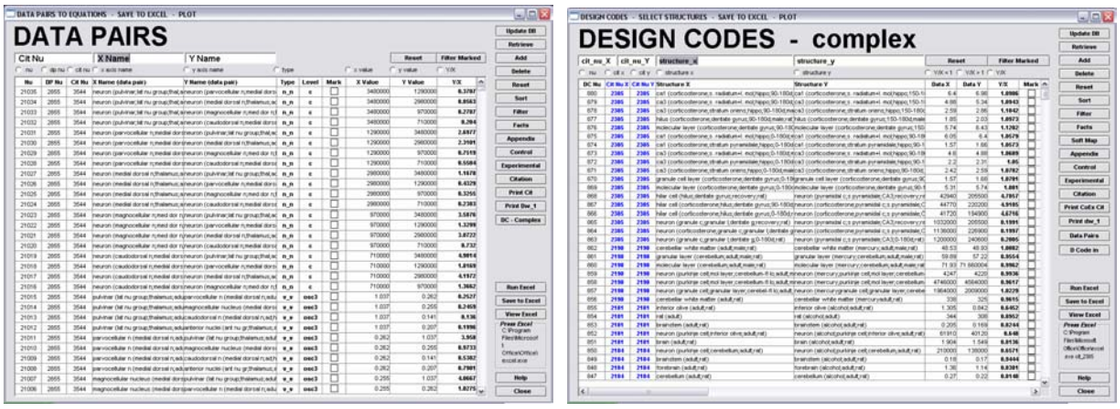

Figure 2: Examples of screens used for selecting data to be analyzed in Excel worksheets.

The procedure is straightforward, consisting of the following steps.

1. Select a structure x and all related structures y.
2. Calculate the ratio of data values (y/x).
3. Sort the y/x ratios (ascending).
4. Send the contents of the screen to an Excel worksheet.
5. Plot the x/y data as a regression (power) equation and calculate the R^2.
6. Change the number of rows until the $R^2 \geq 0.999$, removing outliers as needed.
7. Record the results.

Examples of searches and calculations can be called from the viewing screens (7.2, 7.3. 8.1, 8.2) in the BIOLOGYtabs 2003 program.

Design Code Library

Types: The ***design code library*** (BIOLOGYtabs 2003; 8.1; 8.2) includes sets of images showing design codes as regression curves; they typically illustrate change. The library includes two collections.

1. ***Simple Design Codes***: Identify quantitative and qualitative changes one paper at a time, and
2. ***Complex Design Codes***: Identify quantitative and qualitative changes several papers at a time.

Properties and Rules: Design codes offer several features, including some restrictions.

- A design code is expressed as a power equation ($Y=\mathbf{b}X^{\mathbf{a}}$), plotting a set of related X and Y values (data pairs) and carrying an $R^2 \geq 0.999$. Recall that **b** is the y intercept and **a** the slope. A design code is interpreted by inspecting the values of **a** and **b**. When **a** is close to one, the curve is parallel to the reference line. When two or more curves are more or less parallel (**a**≈1.0), a **b** >1 indicates and increase and a **b**<1 a decrease.

- A plot of X vs. X - control vs. control - serves as a reference line (X=Y; R^2=1.0), representing no change (Figure 3). (N.B., one design code can serve as the reference of another.)

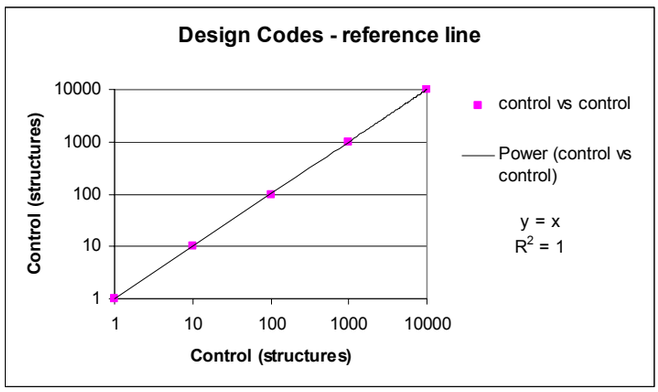

Figure 3. Reference line for design codes.

- A design code equation relies on the reference line for its interpretation. It can be parallel to the reference line (**qual**itatively similar), nonparallel (**qual**itatively different), above the reference line (**quant**itatively more), below (**quant**itatively less), and superimposed (**qual**itatively and **quant**itatively the same). A ***qualitative change*** signifies a new design code wherein the proportion of the parts is different (Figure 4). In contrast, a ***quantitative change*** identifies more or less material in the same proportions (Figure 5). A change often includes both qualitative and quantitative elements (Figure 4).

  - A qualitative change

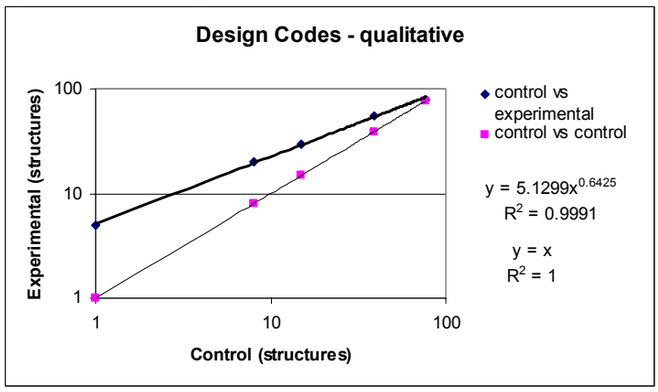

Figure 4. A qualitative change is indicated when the experimental line is not parallel to the reference line (N.B., the experimental line shown above includes both qualitative and quantitative changes).

  - A quantitative change

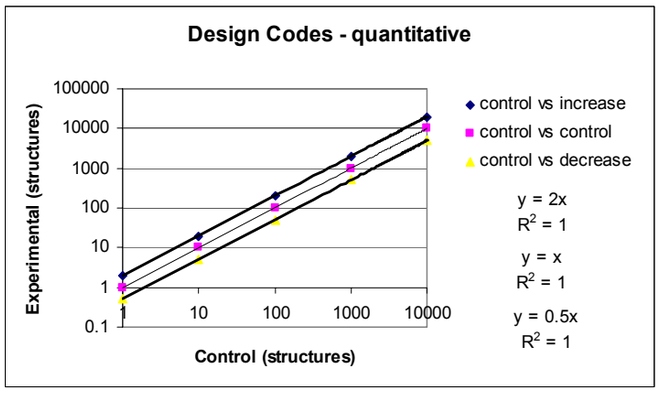

Figure 5. A quantitative change is indicated when the experimental line is parallel to and above or below the reference line.

- A ***complex design code*** (BIOLOGYtabs 2003; 8.2) combines data from several simple design code equations. It characterizes change by structure and by event. 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 only ***qualitative*** changes.

Applications: Tab 8 of BIOLOGYtabs 2003 presents the design codes by topics, each having a separate tab (Figure 6). The collection includes design codes calculated ***by paper*** (8.1: control, experimental, development, aging, disease, structure to function) and ***by papers*** (8.2: complex).

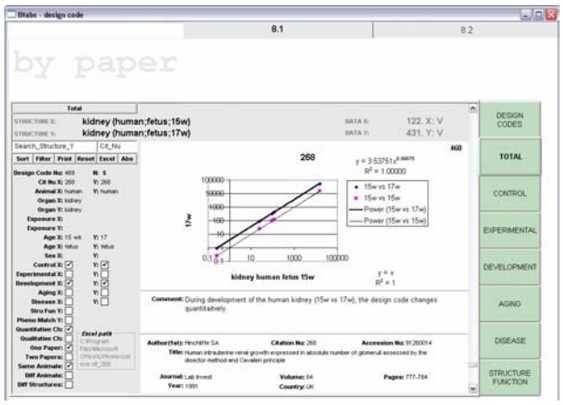

Figure 6. Simple design codes calculated with data from a single paper.

The reader might begin by scrolling through the total list and noticing that change in biology consists largely of quantitative events (as suggested by the prevalence of nearly parallel curves). Major qualitative events will be found largely in the development and experimental tables, but such change tends to be temporary rather than permanent. In development, for example, the "adult" design code is established early in life with subsequent growth characterized largely by more or less parallel curves (***quantitative*** change).

A design code screen offers several ways of accessing the data and graphs. Use the drop down data window (the button is in the upper left hand corner of the screen) to scroll through a list of y axis structures and make selections by clicking on a highlighted line. Alternatively, type in a key word (or the first few letters thereof) into the field labeled "Search_Structure_Y" and press Enter. For example, <sch> will retrieve all the schizophrenia graphs and <hu> all those from the human. To view data from a specific paper, type in the citation number and press Enter. More advanced searches can be run using the sort and filter buttons (for examples of scripts, see Bolender, 2002). To view the scratch sheets that were used to make the graph - currently shown on the screen - click on the citation number and then on the Excel button. Click on the Abs (abstract) button to read an abstract of the paper online. A help file containing more information can be called from the top page of the folder. The data entry screens used to populate the design code tables can be found in the appendix of the main program (EBS 1.0; 2001) – after the 2003 upgrade has been completed.

Tab 8.2 of BIOLOGYtabs 2003 includes examples of complex design codes (figure 7). Notice that the results are often grouped according to the data pair ratios. The ratio (Y/X) is reported as a value less than (<1) or greater than (>1) one, where <1 identifies a decrease and >1 an increase. The histogram, which shows the distribution of these Y/X data, identifies the extent to which a data pair can differ. The lung and liver show considerable change, for example, whereas the brain shows relatively little. Note too that most change appears quantitative – not qualitative, as indicated by slopes with values close to 1.0.

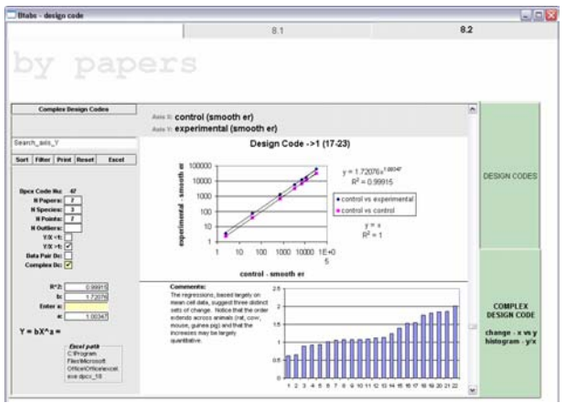

Figure 7. Complex design codes calculated with data from several papers.

Ladder Equation Library

Types: The ***ladder equation library*** includes a collection of ladder (exponential) and rung (power) equations that together summarize data pairs for total and selected data sets. The summary takes the form of a single exponential equation: $y=e^{xa}$. Usually, this type of equation is used to describe data of the physical sciences (concentrations, radioactive decay, half life, etc.).

Properties: The ladder equation is remarkable in that it can summarize all the structures in the library database - expressed as data pairs – with the single expression:

$y = 0.000134e^{0.7498x}$,

where y equals the y intercept of the power (rung) equations and x the number of the rung (e.g., 1 to 24). Figure 8 illustrates the ladder equation for the total data set.

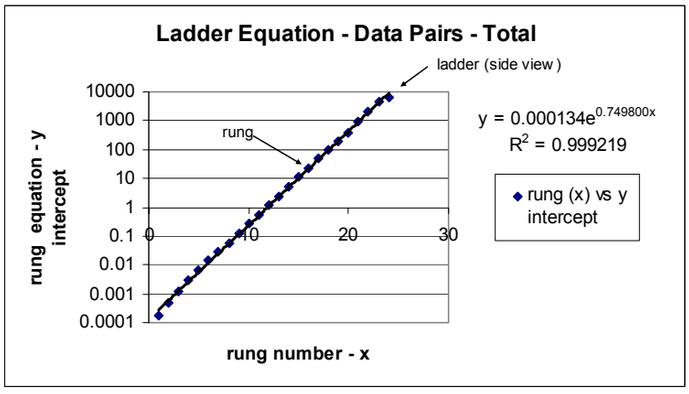

Figure 8. The ladder equation identifies the order of data pairs as an exponential expression.

Ladder equations display several properties.

- They can trace the origin of nonparallel curves in data pairs and design codes to data existing on different rungs or to data moving up or down the rungs of the ladder.
- Data taken from one rung of the ladder tend to produce a power curve parallel to the reference curve.
- The y intercept of a rung equation is twice as large as the rung below and one half as large as the rung above. In other words, moving from one rung to another suggests a quantum (unit) difference. There is either twice as much or half as much. When the requirements of a **critical data set** are satisfied, the amount of change attributed to the structures x and y can be determined. In effect, we will need to know the ratios (Y/X) that change, those that do not, and the absolute amounts of each structure X and Y. Can you imagine where this type of sleuthing might take us?
- Structures - all across the biological hierarchy - can be summarized explicitly by a set of connected equations.

- The rung equations can mitigate the effects of data clustering in controls, wherein points tend to cluster about a point rather than distribute along a line.
- Sets of data characterizing a given structure tend to be positioned on the rungs in a similar order. However, these similar data sets can assume different positions on the ladder (see Table 2). Table 2 illustrates the relationship of the nucleus to organelles – across several species. If these animals have similar genomes, how do we explain these results?

Table 2 Leydig Cell Rungs (nucleus vs. structure i)

Rung	Cit 344 Human	Cit 1403 Human	Cit 220 Mouse	Cit 1365 Mouse	Cit 1208 Guinea Pig	Cit 2405 Rat
2	Golgi		Multivesicular body			
3					Multivesicular body	
4				Multivesicular body		
5			Golgi Peroxisome	Golgi	Peroxisome	
6	Reinke Crystal		Lysosome		Golgi Lysosome	
7	Lipid Droplet					
8		Reinke crystal				Ribosome
9	Lipofuscin	Mitochondrion Lysosome	Mitochondrion Lipid Droplet	Lipid Droplet	Mitochondrion Lipid Droplet	Peroxisome Lipid Droplet
10	Mitochondrion		Cytoplasmic Matrix	Mitochondrion		Lysosome
11						Golgi
12					Cytoplasmic Matrix	
13	Cytoplasm	Cytoplasm		Cytoplasm	Cytoplasm	Mitochondrion
14						Cytoplasmic Matrix

- Rung equations display order as a set of parallel regression curves, having an R^2=0.999 (see Figure 9). Such order seems to be a general phenomenon in biology in that it appears in organs, cells, organelles, etc. Rung equations can tell us something about how structures are constructed and how they change. For example, a quantitative change can be explained as the movement of a data pair from one rung to another– or as no movement at all. If you understand why, then you can use this ***ladder paradox*** to explore yet another level of biological complexity.

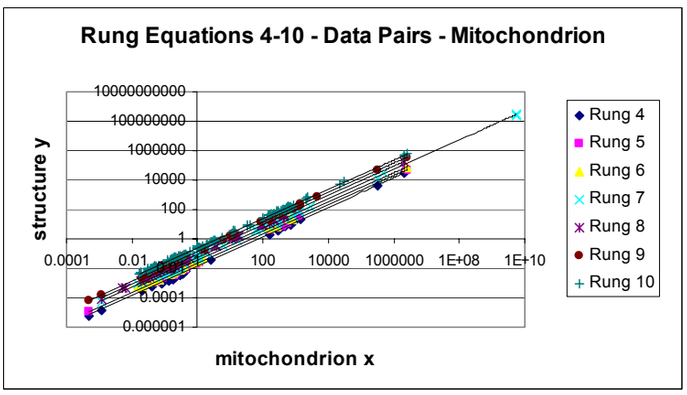

Figure 9. Rung equations for mitochondrion.

Discussion

Biology and Information Science

How does one explore biology as an information science? A curious, yet reassuring pattern that seems to be emerging from the project is that the process of discovery in biology resembles dynamical systems, as described in chaos theory (Waldrop, 1992):

Order → Complexity → Chaos.

However, the ***Enterprise Biology Software Project*** reverses the directions of the arrows. Starting with research data in chaos (scientific journals stored on library shelves), complexity reappears by entering research data into a relational database, and order emerges as equations in derived data libraries. In time, this order may lead us to the laws of nature.

Chaos → Complexity → Order → Physical Laws

Dimensions of Information

Biological stereology is a first rate stepping stone into information science. It allows us to access research data reliably and move them from one dimension to another. These dimensional shifts are basic to reliable data interpretations. Working in different dimensions, however, can be problematic. Recall from the following text from the ***Introduction to Dimensions*** given in Chapter 2 of ***Data City – A Short Story (Bolender, 2001a)***:

"Our familiar world looks very different when viewed from within a given dimension or from different dimensions. In 0 space (space is being used here as a synonym for dimension) you can see points, but not lines, planes, or volumes. In 1 space you can see points and lines, but not planes or volumes, etc. Each dimension therefore has its own set of rules for viewing and interpreting data. Notice that as we move to a higher dimension, the information space becomes enormously richer than the previous one. Grasp the abject poverty of 0 space in contrast to 3 space and it becomes far easier to imagine the astonishing richness of n-dimensional space – especially as n becomes greater than 3"

While these comments where originally directed at biological data, they can be applied as well to the derived data of an information system. Imagine information as a platform from which we can view the same stereological data from different dimensions (0, 1, 2, 3, …, n). A 0 dimension view would be expected to suffer from "abject poverty," whereas each higher dimension would reveal a greater wealth of information. Does this mean that we can advance our understanding of biology by merely moving our information viewing platform to a higher dimension?

Yes, of course, but first we need to take a hard look at research biology – as it relates to dimensions. Recall the rule given earlier (Bolender, 2001a):

"Always interpret data of a given dimension with rules appropriate to that dimension."

From an information standpoint, the published research data of experimental biology are being interpreted largely from a platform of 0 dimensions – the familiar change model (Figure 10). To wit: Does the structure or function change, yes or no?

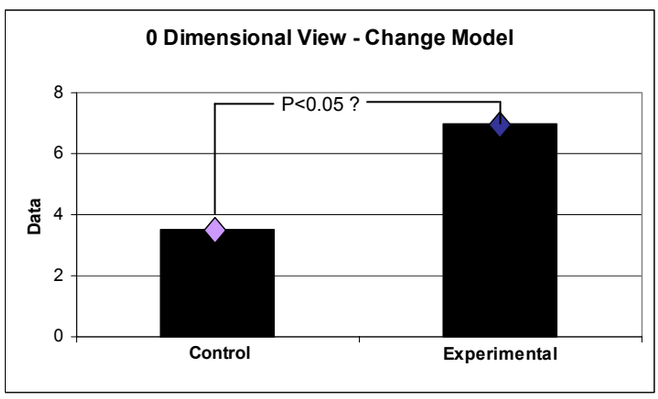

Figure 10. Standard interpretation of research data.

The limitation of this model becomes quickly apparent when one wishes to explain these events in terms of gene function. Once again we face the problem of complexity in that

more than one explanation exists for these results. In a genetically controlled system, we now know that the change shown in figure 10 (experimental) could have been produced by at least three different events (quantitative, qualitative, or quantitative and qualitative).

We also know that explaining gene function will require dimensions higher than zero, because gene function is not an isolated event, but rather one that is highly connected. In fact, the data of a change model is simply a highly restricted view of a connection model (compare Figures 11 and 12). Although most of our experimental papers contain connected data, we have learned to "pull a curtain" around each pair of control and experimental points and then look for a significant difference (Figure 11). Indeed, actively looking for types of change (qualitative; quantitative) and connections within and across data sets is not a common practice in biological stereology.

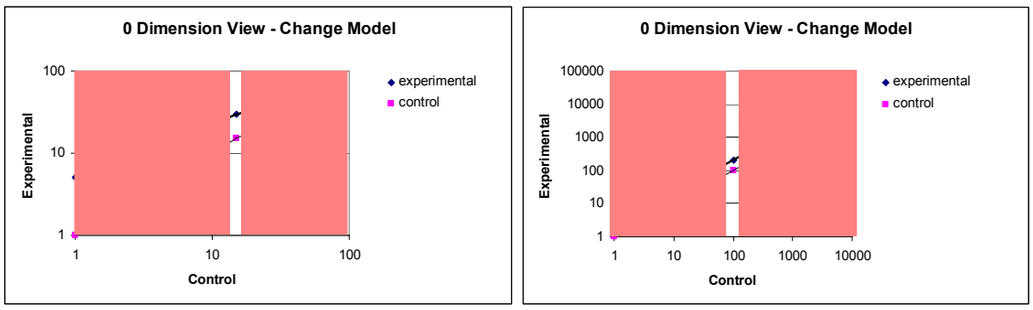

Figure 11. Change Model: Does it change – yes or no?

The consequence of our selecting a 0 dimensional platform is that we as biologists often end up trying to interpret events in dimensions to which we have no access. We forget that only 0 dimensional questions and answers can be explored from a 0 dimension platform. In truth, the price we are paying for maintaining the antiquated change model in research biology is an "abject poverty" of information. To be sure, there is something sadly amiss when the best solution we can find to the problem of complexity in biology is to simply ignore it.

Design Codes

What happens when we move our information platform to a higher dimension? The ***design code*** library allows us to view zero dimensional data (points) from a one dimensional platform (lines). The platform is one dimensional because the library consists of lines that describe how a connected set of structures (and/or functions) are related and how they change. Each design code is produced by fitting 0 dimensional data points to a one dimensional line by calculating a regression equation.

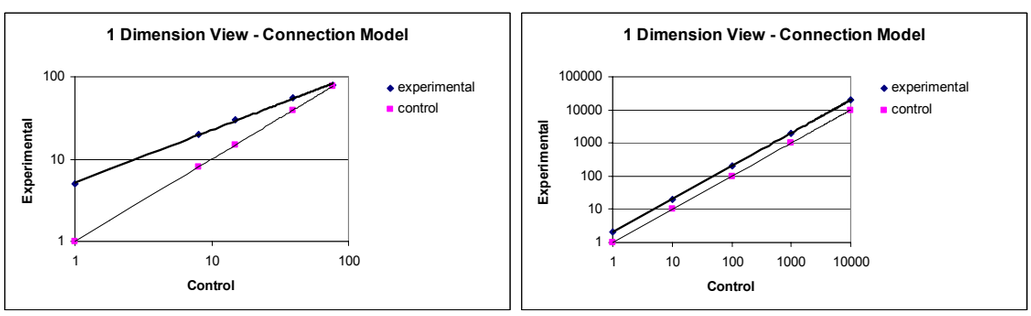

Figure 12. Connection Model: How does it change?

Notice what happens. Once we position ourselves in a dimension higher than zero (Figure 12), we can look back at the 0 dimension and see things that were undetectable from the 0 dimensional platform (Figure 11). For example, we can now clearly see that change – as mediated by a genome – has two distinct and separable properties – y intercept and slope. Recall that $Y = bX^a$, where **b** equals the Y intercept and **a** the slope.

Design codes, which represent a set of connected data pairs, display three types of change: quantitative, qualitative, and quantitative + qualitative. However, data sets can display points beyond the domain of the design code – the outliers. Nuclei, rough

endoplasmic reticulum (RER), lipid droplets, Sertoli cells, and vessels often fall into this category. Currently, the presence of outliers remains unexplained.

Calculating design codes is an exercise in Excel (Figure 13). Data are plotted as regression curves and filtered by eliminating outliers until the power curve displays an $R^2 \geq 0.999$. These worksheets can be called from the design code screens (e.g., see figure 6).

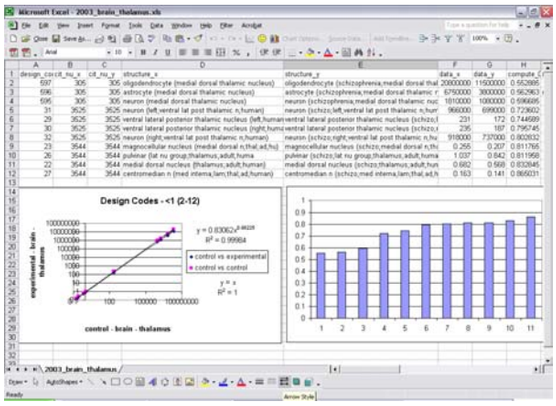

Figure 13. Excel work sheet for design codes.

Design Codes – by paper (simple)

In the software, the design codes are grouped according to the topics listed below.

Control: The design codes are well-suited to asking questions about the principles underlying the design of organs, tissues, and cells. What structures display similar codes and of those structures what parts are similar? Do all exocrine organs share a common structural plan? If so, what differentiates one from the other – in terms of phenotype? What animals and what parts thereof share similar design codes?

Experimental: Typically, biology establishes the *qualitative* design code of an adult early in development of an organ and then maintains the code assiduously throughout life. Indeed, most of the *qualitative* changes in design codes occur early in development and to a lesser degree during disease, aging, and exposure to xenobiotics. Of course, short term changes occur routinely during cycles (e.g., digestive, diurnal, reproductive, etc.).

Experimentally induced changes in design codes (*qualitative* and *quantitative*) often require severe measures. High doses, long treatments, ablations, and extreme conditions are not uncommon to protocols inducing change. In general, maintaining more or less constant design codes throughout life appears as a consistent theme in biology. Genes support a phenotype that defines a normal, healthy individual. Experimental conditions disrupt this normal state.

Design codes also offer the promise of providing effective solutions to otherwise troublesome problems. For example, detecting change unambiguously in biopsy material is problematic because the sample fails to meet all the requirements of a critical data set. Recall that a biopsy is a sample of a larger structure, the volume of which remains unknown. However, in a representative sample, a design code can detect a *qualitative* change unambiguously. Moreover, one can infer that a *quantitative* change in an average cell is proportional to a change in a parent structure – when the number of cells remains constant (see the mathematics course; Bolender, 2001b). To support the validity of such an approach, design codes from biopsies could be compared with those coming from organs estimated with unbiased sampling methods. It is of interest to note that Bauer et al. (2001) have already used ratios successfully for identifying patterns among cells in the skin.

Development: After viewing this collection of design codes, one is left with the impression that development follows a general pattern. It consists first of establishing

structures – and their parts - according to an adult design code and then growing larger. That is to say, *qualitative* change is soon replaced largely by *quantitative* change. Notice as well that the equations pinpoint the stage of development when *qualitative* changes in the design code occur. Such stages, when more of the genome may be involved, might correspond to periods of heightened susceptibility to damage from external factors, such as malnutrition or xenobiotics.

Aging: In comparison to development, aging often resembles negative growth. The design codes suggest that aging typically reflects a loss in structure (*quantitative* change) rather than a rearrangement of parts (*qualitative* change). If this pattern persists, then those factors that encourage a *quantitative* increase in structures, such as hormones, may one day be used to mitigate the aging process. Of course, the factors responsible for aging are likely to include more than just a progressive loss of structure. In contrast, for example, connective tissue compartments can show an increase with aging.

Disease: A disease, its development, and remission can be described quantitatively and qualitatively with design code equations. In addition, the equations simplify the task of comparing results across studies and encourage the search for patterns of structure and function within and across diseases. Again, the ability to distinguish between quantitative and qualitative events – presumably being initiated by the genome - may have important consequents in developing strategies for treatment and prevention.

A disease can be summarized explicitly as a function of variables, using one or more design code equations. The variables may be structural, functional, or both. This offers a convenient approach to pathology because it defines disease with equations having both qualitative and quantitative characteristics. In designing treatment protocols, for example, it will be useful to know at the outset if more or less of something is needed or if the proportion of the parts needs to be adjusted. In either case, the success of a treatment strategy might be improved importantly by knowing the design code(s) first.

Relationships of Structure to Function: One of the central goals of modern biology is to understand relationships of structure to function. In a 0 dimensional model, such relationships have relied almost exclusively on showing ***correlations*** between changes in structures that paralleled changes in biochemistry or physiology. Such comparisons are indirect. In contrast, the design code with its extra dimension can express relationships of structure to function directly by fitting structural and functional variables to the same regression equation. Such an equation – or set of equations - creates a direct mathematical pathway to and from the genome, spanning all levels of the biological hierarchy of size. By establishing strong links between structure and function, design codes may offer new opportunities – in collaboration with the data of other disciplines - for developing large scale systems for diagnosis and prediction in research and health care.

Design Codes – by papers (complex)

Complex Design Codes: To look for larger patterns in biology, we can combine the design code equations from several papers to produce summary equations. This offers us a simple way of seeing the bigger picture. Let's look at some examples.

A plot of the total design code data (control vs. experimental) produces a regression curve with an R^2 of 0.98381 for all organs from all animals, as shown in Figure 14. What does this mean? Recall that the design codes detect the amount and type of change that occurs between a set of control and experimental points. The plot below suggests that change occurs by rule.

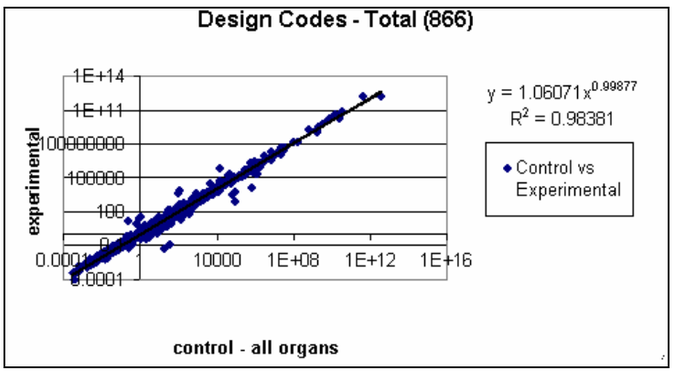

Figure 14. Plot of all control design codes vs. all experimental design codes.

Now, if we plot just the design code data for the brain (50 points) from all animals and all experimental changes, then we find a completely unexpected result (Figure 15). Now practically all the points fall on the regression line and the R^2 is 0.9992. What does this mean? It tells us that the brain regulates change very tightly. The big surprise was the R^2 so close to one, considering the heterogeneity of the experimental settings. ***In effect, change behaves as a highly predictable event.*** That is to say, the data from future experiments on the brain would be expected to fall on the line $y = 0.9773x^{0.9224}$. Although such a statement appears highly speculative, none-the-less, the pattern persists (see Table 3).

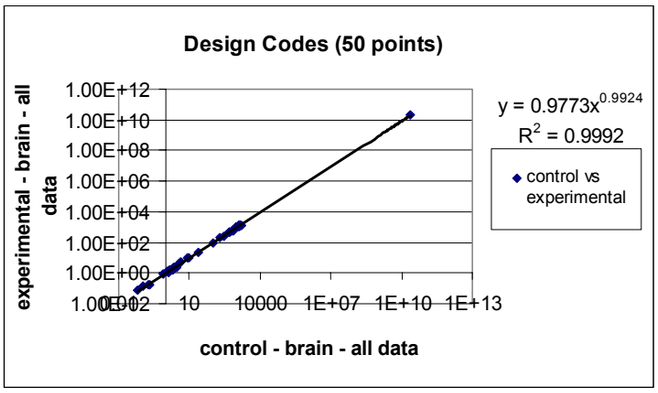

Figure 15. Plot of all brain design codes: control vs. experimental .

This unusual view of change in biology becomes even more intriguing when one browses through the complex codes in tab 8.2 of BIOLOGYtabs 2003. The pattern seen for the brain is repeated time and time again for other organs. Change behaves as a rule-based event. In the table below, each structure or event changes according to the rules defined by its equation(s) – apparently ***independent of the experimental setting***. For this to make sense, it may be helpful to recall that design codes plot just the amount of change - not the actual values.

Table 3. Predicting Change. Y (experimental value) = X (control value) × b[a]

BY STRUCTURE	*b*	*a*	*R^2*
Brain Cerebellum	1.11242	0.98779	0.99975
Brain Nuclei	0.84944	0.99143	0.99935
Brain Thalamus	0.83062	0.98225	0.99984
Kidney Glomerulus	1.04964	0.98910	0.99942
Lung (Y/X<1)	0.84280	0.98491	0.99897
Lung (Y/X>1)	2.63090	1.02137	0.99907
Lysosome	1.28034	0.98597	0.99898
Nucleus (Y/X<1)	0.76493	0.99727	0.99920
Nucleus (Y/X≥1)	1.06734	1.02496	0.99995
Parathyroid	2.01231	0.99392	0.99928

BY EVENT	*b*	*a*	R^2
Aging (Y/X<1)	0.82531	1.00899	0.99966
Aging (Y/X>1)	1.05548	1.00422	0.99916
Alcohol (Y/X<1)	0.85233	0.99278	0.99969
Alcohol (Y/X>1)	1.39447	1.09812	0.99982
Dementia (Y/X<1)	0.73290	1.01268	0.99905
Dementia (Y/X>1)	1.08507	0.99604	0.99996
Malnourished	0.88698	0.98397	0.99948
Schizophrenia (Y/X<1)	0.87118	1.00651	0.99972
Schizophrenia (Y/X>1)	1.15992	0.98826	0.99945

Complex design codes offer several additional features.

Predicting results
In BIOLOGYtabs 2003 8.2, each screen includes a working version of the regression equation that can be used to predict Y from X. Enter a value into the data entry field "enter x" and press **Enter**. In effect, these regression equations should predict the results of many experiments yet undone.

Summarizing data across labs
One of the pleasant surprises to come from this new library is that it allows us to combine data coming from several different laboratories by merely fitting similar data to the same regression curve(s). This gives us a way of characterizing structures with equations - across animals, organs, and events.

Looking for stereological evidence that gene sequences are conserved across species
We know from sequence data that genes can be remarkably similar from one type of animal to another (Waterston R. H., et al., 2002). We also know that that a genetic sequence on a chromosome and its ultimate expression as a downstream structure – with a recognizable function - may be influenced by a variety of intervening factors. In other words, a similarity in genotype may not always produce a corresponding similarity in phenotype. Design codes might offer a convenient way of exploring similarities and differences in phenotypes derived from a range of genotypes.

Ladder Equations

Ladder and rung equations may help to explain an unexpected source of variability in research data. Consider what these equations might be telling us. If the process of making of a gene product is optimized, then producing twice as much product would require the activation of second gene. Similarly, to decrease production by one half, one half of the active genes would be turned off. If so, then the appearance of gene products in cells and tissues might be expected to occur in a similar way. In other words, the distribution of biological parts would naturally fit an exponential (ladder) equation, wherein each rung might represent a different level of gene activation.

We know from DNA sequencing that genomes contain considerable amounts of redundant DNA – multiple copies of the same genes or of genes that produce similar products (Nowak et al., 1997; Young et al., 2003). Perhaps it is exactly this multiplicity of DNA that allows for change in biology and expands diversity within and across species (see Table 2).

Redundancy would seem to be a practical solution for a control system based on a switching mechanism. Consider the liver. It uses "redundant DNA" to meet its productivity requirements by having polyploidy nuclei (e.g., 4n, 8n). A hepatocyte can readily double or quadruple its output by simply switching on its extra copies of DNA. Moreover, the "switching" can also include a temporary transformation from a mononucleated to a binucleated cell. The advantage of ladder equations is that they offer a convenient way of studying change as a stepwise (quantum) event.

Stoichiometry

In biology, order can be found everywhere one looks – if one looks in the right places. Using stereological data, we can see this order routinely with ***data pairs***, ***design codes***, and ***ladder equations***. Moreover, the order is consistent, scalable, and predictable. The question before us, of course, is whether stereology can help us to find the source of all this order.

If order springs from underlying order, then perhaps biology solved its problem of assembling structures from molecules, organelles, cells, organs, etc. by leveraging the stoichiometric order already present in physical systems. Indeed this would have been an elegant solution. Since stereology offers direct connections between structure and function, it offers ample opportunities for testing such a hypothesis.

Caveats

We can think of the ***Enterprise Biology Software Project*** as a "three year old" trying to understand complexity in biology. As a data-based approach, it relies on generous amounts of data for finding patterns. The data pairs library has grown to 20,000 entries, but those of the design codes (518) and ladder equations (63) are still quite small. More over, data that do not fit the R^2=0.999 curves – and outliers – have been largely ignored.

Concluding Comments

What have we learned from the ***Enterprise Biology Software Project*** that we didn't know at the start? When we elect to explore biology as an information system, research data become amenable to a robust quantitative approach provided we standardize them and minimize bias. We know that a connection model (1D) provides more information than a change model (0D) and that real-world data can be expressed with coefficients of variation very close to one ($R^2 \geq 0.999$). These new data appear not as a rare curiosity, but as a widespread event. We can also identify a strategy for dealing with complexity, which consists of constructing libraries and platforms for addressing specific problems. This approach becomes evermore inviting in that each new library creates abundant opportunities for discovery – quite often unimagined at the outset.

By taking a page from the book of theoretical physics, we have seen that intractable problems can be tackled convincingly by moving our viewing platform to a higher dimension. Shift our view of stereological data from a 0D model (change) to a 1D model (connection) and the complexity of change unfolds neatly into two distinct parts – qualitative and quantitative. The observation is an important one because it illustrates the difference between looking at our results in and out of context – the principal difference between connectionism and reductionism.

References

Bauer J, Bahmer FA, Worl J, Neuhuber W, Schuler G, Fartasch M. 2001 A strikingly constant ratio exists between Langerhans cells and other epidermal cells in human skin. A stereologic study using the optical disector method and the confocal laser scanning microscope. J Invest Dermatol 116:313-318.

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

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997 Evolution of genetic redundancy. Nature 388(6638):167-71.

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

Waterston R. H., et. al. 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520-562

Young TL, Xiaoshan D, Guo D,King RA,Johnson JM,Rada JA. 2003 Identification of genes expressed in a human scleral cDNA library. Mol Vis 9:508-514.