In: Enterprise Biology Software, Version 7.0 © 2007 Robert P. Bolender

Enterprise Biology Software: VIII. Research (2007)

ROBERT P. BOLENDER

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

Biological data differ remarkably from those of physics and chemistry in that their properties are intimately tied to their physical locations. A hierarchical arrangement - consisting of parts contained within parts - defines biology as a complex set of interacting complexities, scaling in size from molecules to organisms. None of this uniqueness becomes apparent until we start putting the parts back together. If, for example, we don't return the parts to their original *in vivo* settings before attempting an interpretation, we run the risk of allowing our data to become meaningless or at best severely limited. In effect, a failure to recognize this special property of biology leads inevitably to a semiquantitative science. Our task this year therefore becomes one of understanding how biology can become broken experimentally and then showing how we can fix it. To do this, we will explore ways of extending what we have learned from the stereology literature to disciplines often reduced to relying on semiquantitative approaches, such as biochemistry and molecular biology. We begin by defining the boundary conditions for experimentation in biology, explore the unstable foundations of semiquantitative data, and finally suggest workable alternatives.

Specifically, here is our question. Can we leverage our ability to create a mathematical biology for the first fourteen levels of the biological hierarchy to the remaining two, namely molecules and genes? To answer this question convincingly, we must identify a general solution to the problem of counting molecules in an *in vivo* setting for biochemistry, molecular biology, and immunocytochemistry. This gives us an interesting, but somewhat difficult puzzle to solve because a solution depends on solving several smaller puzzles first. This tendency of having to solve intricate puzzles embedded in intricate puzzles offers a hint of what biology has in store for us. In any case, the general solution to the counting molecules problem becomes the ***hybrid hierarchy equation*** – one that employs gold standards to work its magic. The main product to emerge from the effort this year is a rule book, one that carefully addresses the uniqueness of biology as a scientific discipline. The software package for 2007/8 (Enterprise Biology Software, Version 7.0) includes new data harvested largely from years 2004-5, expands the biology blueprint, adds cluster analysis, launches the hybrid hierarchy equation, and offers guidelines for a mathematical biology.

Introduction

Transforming biology into a quantitative science begins by changing the way we manage our research data. An essential first step consists of moving published data from the pages of journal articles into the tables of relational databases where they can be standardized and used to look for mathematical patterns. Such patterns can be readily found in the connections that occur between parts and captured as data pairs and repertoire equations. In turn, these data pairs and equations can become the raw material for building a mathematical biology. Fundamental to the success of this approach includes an ability to manage complexity in biology by mathematically unfolding and refolding structural relationships - throughout the biological hierarchy of size. By leveraging the mathematical order inherent in biological systems, we now have a workable strategy for reverse and forward engineering structures of all sizes – one based largely on our ability to tap into the mathematical core of biology. From this we learn that mathematics and technology can become powerful discovery tools when we use them in harmony with the intrinsic order of biology. All of this becomes possible when we transform biology into a database science.

THE MODELS

Models define the ways in which we explore, interpret, and discover biology. Two models, reductionism and change, largely define the research biology of today and have been eminently successful in producing largely a descriptive and semiquantitative biology. However, the major limitation of these models is their inability to address biology as a complex adaptive system. One approach to overcoming this limitation is to upgrade our current models and add new ones. To make this transition, however, we will need a mathematical stepping-stone to take us from semiquantitative to quantitative. Stereology can serve remarkably well in this role and connection, integration, and engineering models offer new discovery platforms that can operate comfortably in a complex setting.

Reductionism

Reductionism reduces or eliminates complexity by extracting specific parts of biology so that they can be carefully analyzed and studied. The extraction process, however, usually forfeits the structural order of the part, severs connections with other parts, and produces isolated data. When such data are interpreted out of context, they often become semiquantitative. In other words, perfectly good quantitative data can collapse into a semiquantitative state by simply assuming they can do something they are incapable of doing.

Change

The change model compares two data points – usually a control and experimental – with the goal of detecting a significant change. Detecting a significant change in isolated data is extremely easy to do in a semiquantitative setting, whereas detecting a significant change – that is also valid - in a biological setting requires the power of a quantitative biology. Fortunately, the change model can be readily upgraded to quantitative by merely designing experiments as equations that include all the required variables.

Connectionism

The connection model maintains that all parts of biology are connected by rule, which means that the structural order of biology can be captured with equations. Here the practical solution consists of standardizing published research data by moving them into a relational database and then forming data pairs. In turn, these data pairs can be assembled into repertoire equations that define the connections mathematically. This model offers a robust solution to the problem of unfolding and refolding complexity.

Integration

The integration model operates by storing standardized data of all biological disciplines in a single database table and then using them to form data pairs and decimal repertoire equations. This cleaning and summarizing process leads to a universal biology database. Such a database effectively integrates published research data across disciplines and experimental settings.

Engineering

The engineering model uses biological data to unfold (reverse engineer) and refold (forward engineer) biological complexity. This can be done locally with hierarchy equations and locally and globally with decimal repertoire equations. This model established a quantitative foundation for diagnosis and prediction.

MATHEMATICAL BIOLOGY – A Brief Introduction

The mathematical biology described herein depends wholly on published data to generate the empirical equations that allow us to explore an information space of uncommon complexity. To qualify, these data must be gathered with unbiased sampling methods within the framework of a rule-based approach. The brief introduction that follows considers three basic components of a mathematical biology that will serve to illustrate how this quantitative approach to biology works. The **Rule Book: Guidelines to a Mathematical Biology** continues this process, but in greater detail (Bolender, 2007).

Unbiased Sampling

In biology, sampling is everything. Unbiased sampling requires that all parts of the structure being sampled must have an equal chance of being sampled. Any other sampling scheme automatically becomes suspect. Samples collected for biochemical analyses that do not come from a total cell or tissue homogenate, or come from isolated cell or tissue fractions will also fail this test – unless the data are collected and interpreted within the framework of analytical fractionation (de Duve, 1974).

Experimental Design

Research experiments can become consistent with the organizing principles of biology when they are designed as hierarchy equations. The process is surprisingly straightforward. The equations define the problem in terms of variables, which, in turn, are collected as data in the laboratory by applying unbiased sampling methods. Finding a solution to an experiment consists of entering the experimental data into an equation and evaluating it. The challenge for the reader will be to learn how to write balanced hierarchy equations. However, the reward for such an effort can be substantial. This skill will save countless hours in designing experiments, writing discussions, reading research papers, and reviewing manuscripts and research proposals.

Biological Data

Data appearing in the biology literature can be quantitative, semiquantitative, or descriptive. To qualify as quantitative data in a mathematical biology, they must be clearly identified, satisfy the unbiased sampling requirement, and detect differences and changes unambiguously. Being quantitative also depends on how, where, and when the data are being used. For example, quantitative data in one setting can quickly become semiquantitative in another.

Let's begin. Most biological data represent directly or are derived from four basic quantities: volume (V), surface (S), length (L), and number (N). Recall that weight includes the product of a volume (V) and a density (ρ): $W=V \times \rho$.

In biology, however, these four basic quantities can become linked mathematically in curious ways because of the hierarchical organization of the parts. When parts are contained within parts, with cells serving as the basic unit of construction, dependencies occur that become mathematically inseparable. This means that the parts cannot be separated from one another when they belong to a quantitative unit. Let's look at an example.

Biological parts arranged in a structural hierarchy become a function of three variables: volume (V), mean volume (meanV), and number (N). This gives three relationships.

$V = \text{mean}V \cdot N$
$N = V / \text{mean}V$

Mean V = V / N

If we want to know at any given time what biology and its parts are up to, we need to know what's happening to these three variables. Notice that in practice we have to measure (or estimate) only two of the variables, because the equation allows us to solve for the third. What happens when we decide to separate these three variables in our experimental design? We get exactly what we don't want, namely an incomplete or semiquantitative result. In effect, by breaking them up, we break the rules. Reductionism tells us we can, but we really cannot. Why? ***We can take the data out of the hierarchy, but not the hierarchy out of the data.*** This turns out to be a fundamental property of biology, behaving as a complex hierarchical system. It is a central principle of experimental biology widely unknown yet enormously important.

Now, let's work through a few examples. We begin with the general equation for volume, wherein a compartmental volume is the product of the mean volume of a part and the number of parts:

$V = \text{mean}V \cdot N.$

By adding subscripts for a cell (cell), we can focus on the behavior of a specific part, namely a cell:

$V_{\text{cell}} = \text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})}$, where for convenience we will assign centimeter units:

$\text{cm}^3 = \text{cm}^3 \cdot \text{cm}^0$; recall that $\text{cm}^0 = 1$.

This equation tells us that to get the total volume of a compartment of cells, we need to know both the mean cell volume and the cell number. Alternatively, we can get the same information by knowing the concentration ($V_{\text{cell}} / V_{\text{structure}}$) of the cells in the containing structure ($V_{\text{structure}}$):

$V_{\text{cell}} = V_{\text{structure}} \cdot (V_{\text{cell}} / V_{\text{structure}})$, where $\text{cm}^3 = \text{cm}^3 \cdot (\text{cm}^3 / \text{cm}^3)$.

Now, lets compare the information content of these two equations.

(1) $V_{\text{cell}} = \text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})}$
(2) $V_{\text{cell}} = V_{\text{structure}} \cdot (V_{\text{cell}} / V_{\text{structure}})$

Equation (1) contains information about the volumes and numbers of the parts (i.e., cells), whereas equation (2) contains information only about the volumes of the parts. To interpret a change in the cells (V_{cell}) unambiguously, equation (1) will work but not equation (2). Why? Because $V_{\text{cell}} = \text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})}$. A change in the volume of cells can be influenced by a change in the mean cell volume, a change in the number of cells, or a change in some combination of the two. Moreover, a change in equation (2) can be influenced by a change in the volume of the parent structure ($V_{\text{structure}}$) plus all the changes that can occur in equation (1).

By combining equations (1) and (2), we get:

$\text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})} = V_{\text{structure}} \cdot (V_{\text{cell}} / V_{\text{structure}})$.

We can then solve for the concentration of cells ($V_{\text{cell}} / V_{\text{structure}}$):

$(V_{\text{cell}} / V_{\text{structure}}) = (\text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})}) / V_{\text{structure}}$, where $(\text{cm}^3 \cdot \text{cm}^0) / \text{cm}^3 = \text{cm}^0$.

Why is this equation important? It unveils the complexity of a cell concentration. Measuring a concentration is easy, explaining **why** a concentration has changed is considerably more difficult because more variables come into play. Here's the point. When comparing experimental to control concentrations (e.g., cells), we are dealing with at least four variables in the numerators and two in the denominators for a grand total of six variables - all of which can contribute to the outcome. Remember that more than two variables in play effectively renders an experimental result ambiguous (see **Rules #4** and **#5** in the **Rule Book** (Bolender 2007)).

Consider this. When we add counts of molecules to our experiment, the cells and all the containing structures (parts in parts) continue – as variables - to influence the outcome of the experiment. This explains why it's so important to express all experiments as hierarchy equations.

Now let's see what happens when we want to detect changes in molecules. We begin by writing the equation of our experiment, one that will give us the number of molecules in a population of cells contained within a structure ($V_{\text{structure}}$). The same equation applies to control and experimental data. Color-coding identify the variables with units and phenotypes that can cancel.

The usual stereological equation is given as:

$N_{(\text{molecules,cell})} = V_{(\text{structure})} \cdot Vv_{(\text{cell/structure})} \cdot Nv_{(\text{molecules/cell})}$.

Next, we can unfold it to view the inherent complexity:

$N_{(\text{molecules,cell})} = V_{(\text{structure})} \cdot \{(\text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})}) / V_{(\text{structure})}\} \cdot \{N_{(\text{molecules})}) / (\text{mean}V_{(\text{cell})} \cdot N_{(\text{cell})})\}$, where

$\text{cm}^0 = \text{cm}^3 \cdot (\text{cm}^3 \cdot \text{cm}^0 / \text{cm}^3) \cdot (\text{cm}^0 / \text{cm}^3 \cdot \text{cm}^0)$, where $\text{cm}^0 = 1$.

Now we are ready to design an experiment that will let us detect a change in the number of molecules in a biological setting and to explain the source(s) of the change.

In Vivo and In Vitro Experiments

The experiment will compare the number of molecules in the experimental ($N_{\text{molecules(E)}}$) to those of the control ($N_{\text{molecules(C)}}$):

$(N_{molecules(E)}) / (N_{molecules(C)}) =$
$\{V_{structure(E)} \cdot \{(meanV_{cell(E)} \cdot N_{cell(E)}) / V_{structure(E)}\} \cdot \{N_{molecules(E)}) / (meanV_{cell(E)} \cdot N_{cell(E)})\} /$
$\{V_{structure(C)} \cdot \{(meanV_{cell(C)} \cdot N_{cell(C)}) / V_{structure(C)}\} \cdot \{N_{molecules(C)}) / (meanV_{cell(C)} \cdot N_{cell(C)})\}.$

What does this equation tell us? When we wish to count molecules in an *in vivo* setting several variables come into play. Recall that a common way of counting molecules in biochemistry and molecular biology is to measure an optical density (OD) in a homogenate or a sample isolated there from. Such a measure represents a molecular concentration - the number of molecules contained within a unit of containing volume. Now let's ask the telling questions. If we count molecules with optical densities, will our results be the same as those coming from the hierarchy equations given above? Will we be able to explain the changes in terms of real biological events? Will the equation below work in a biological setting? Is the following equation correct?

$(N_{molecule(E)}) / (N_{molecule(C)}) = (OD_{molecule(E)}) / (OD_{molecule(C)}) =$ Correct?

The answers to the four questions above are no, no, no, and no. Why? Because an optical density generates only a molecular concentration, it ignores what's happening to everything else, and therefore produces an ambiguous outcome. In large part, optical densities are responsible for creating an environment wherein the quantitative data of biochemistry and molecular biology are being routinely – and regrettably - downgraded to a semiquantitative status when used to detect changes.

In an experimental setting, an optical density experiment simply compares two molecular concentrations, one control and one experimental. The method assumes that comparing optical densities is equivalent to comparing molecular numbers and that the following equation must correct.

$OD_{molecule(E)} / OD_{molecule(C)} =$
$(N_{molecule(E)} / V_{reference(E)}) / (N_{molecule(C)} / V_{reference(C)}) = N_{molecule(E)} / N_{molecule(C)})$

To accept this assumption, what must be true? For the above equation to work, the ***reference volumes*** and the ***reference phenotypes*** must cancel.

Reference volume: $V_{reference(E)} / V_{reference(C)} = 1$, where, for example, $cm^3 / cm^3 = 1$.

Reference phenotype: $V_{reference\ phenotype(E)} / V_{reference\ phenotype(C)} = 1$, where the contents of the cm^3 reference in the control $(meanV_{cell(C)} \cdot N_{cell(C)})$ is exactly the same as contents of a cm^3 reference in the experimental $(meanV_{cell(E)} \cdot N_{cell(E)})$. Recall that $V_{cell} = meanV_{cell} \cdot N_{cell}$.

If these control and experimental reference volumes fail to cancel, then the optical density equation produces an ambiguous result because it is being influenced by the behavior of four variables:

$OD_{molecule(E)} / OD_{molecule(C)} = (N_{molecule(E)} / V_{reference(E)}) / (N_{molecule(C)} / V_{reference(C)}),$

which can be unfolded to give six variables:

$OD_{molecule(E)} / OD_{molecule(C)} = (N_{molecule(E)} / (meanV_{cell(E)} \cdot N_{cell(E)})) / (N_{molecule(C)} / (meanV_{cell(C)} \cdot N_{cell(C)}))$.

The point being made here is that experiments in biology cannot be well designed in the absence of a realistic and correctly balanced hierarchy equation.

Unrealistic Experiment

$N_{molecule(E)} / N_{molecule(C)} = OD_{molecule(E)} / OD_{molecule(C)}$

Realistic Experiment

$OD_{molecule(E)} / OD_{molecule(C)} = (N_{molecule(E)} / (meanV_{cell(E)} \cdot N_{cell(E)})) / (N_{molecule(C)} / (meanV_{cell(C)} \cdot N_{cell(C)}))$.

If a unit of reference volume is always a standard unit of volume (e.g., one cm^3), why does it fail to cancel? It a biological setting, a cm^3 of reference volume can cancel only when the contents of the cm^3 in the control and experimental settings are exactly the same. What can change in a cm^3? The number of molecules, the number of cells, the number of molecules in the mean cell, the size of the cells, the shapes of the cells, the amount of interstitial material, and a host of changes that can occur in all the other parts that may be present. It turns out that a cm^3 of reference volume represents a ***reference phenotype***, one that will be unique to each control and experimental data point. This means that in such a setting a cm^3 or gram of tissue cannot be expected to cancel. ***Understanding the role of the reference phenotype will allow both biochemistry and molecular biology to prevent their research data from becoming semiquantitative – just as it has already done for the densities of biological stereology***.

Just how important is this reference phenotype rule? It turns out to be ***very*** important. Is there, for example, any convincing way of judging quantitatively the impact of the ***reference phenotypes*** on the results of an experiment? Yes. If we turn to the ***Stereology Literature Database*** and collect data from studies that reported both the number (N) and concentration (N/V) for the same part(s), then we can see firsthand how often the two estimates agree.

To this end, I have updated the percentage change table, relabeled it the ***Concentration Trap,*** and included a fresh copy in the current software package. This software tool can be quite helpful in that it allows us to estimate the amount of damage to expect when concentrations are being compared in an experimental setting – both locally and globally.

Some examples may help. Globally, changes in the same part detected as a total number of parts (N) and as a concentration (N/V) agree only about 50% of the time. Since we can expect a similar outcome for OD data, how might we expect to interpret an experiment based just on optical densities? If the OD data in a paper are reported to be significantly different at the 95% level ($p<0.05$), how should we be interpreting these results? Since such a comparison – on average - enjoys only about a 50:50 chance of being correct, this could reduce the overall significance level of the outcome to a probability of less than 50%. Such an outcome is clearly an ambiguous one. It tells us that concentration (OD) data become unreliable when they are pulled out of the equation of an experiment and used on their own to detect a change. They simply cannot do it alone because they need the help of other variables and the direction that comes from a rule-based equation.

On close inspection, OD comparisons – control vs. experimental – clearly do not carry enough information to detect a biological change reliably. If we can demonstrate convincingly that the concentrations being detected with optical densities suffer the same fate as the concentrations (called densities) of stereology, then OD data coming from ***in vivo*** experiments become suspect and will most likely warrant a similar downgrading to semiquantitative.

Let's look for solutions to this OD problem, because it seems quite unlikely that that these data can survive the downgrade - when put to the test described above. To use OD data productively, we must figure out how to reduce the mischief being caused by the variability of the reference phenotypes. Recall that the reference phenotype of the control data will not cancel the one of the experimental because they are different parts phenotypically. This leaves us with at least two workable options. We can use the approach illustrated by the equations given at the beginning of this discussion, or we can eliminate the reference phenotype locally – at the level of the control and experimental data points. This is accomplished by forming data pairs, each part of which must be related to the same reference phenotype.

Local Fix: Form data pairs to upgrade semiquantitative OD data to quantitative.

When both molecule A and B are related to the same reference phenotype, then the following data pairs can be formed.

For control data

$$OD_{\text{molecule A(C)}} / OD_{\text{molecule B(C)}} = (N_{\text{molecule A(C)}} / V_{\text{reference ph (C)}}) / (N_{\text{molecule B(C)}} / V_{\text{reference ph(C)}}) = N_{\text{molecule A(C)}} / N_{\text{molecule B(C)}}$$

For experimental data

$$OD_{\text{molecule A(E)}} / OD_{\text{molecule B(E)}} = (N_{\text{molecule A(E)}} / V_{\text{reference ph(E)}}) / (N_{\text{molecule B(E)}} / V_{\text{reference ph(E)}}) = N_{\text{molecule A(E)}} / N_{\text{molecule B(E)}}$$

Within the framework of a connection model, these optical density data can now provide reliable information about the proportions of molecules (e.g., A:B) in both control and experimental settings. However, these data provide no information about the total number of molecules. The point to be made here is that semiquantitative OD data can be transformed into a reliable form of quantitative data. Once expressed as data pairs, the OD data can be added to the **Universal Biology Database**, interact with other data types, and contribute to forming equations and uncovering mathematical patterns.

Global Fix: ***Design the experiment as a hierarchy equation and collect all the essential variables.***

Mathematical Phenotypes

The term ***mathematical phenotype*** describes the relationship of biological parts, from two (data pairs) to many (repertoire networks). Here we consider two ways of assembling such phenotypes.

Decimal Repertoire Equations: Parts sharing the same (or adjacent) decimal repertoire equations display stronger associations than with those located in more distant equations. Such information becomes useful when looking for related parts or when assembling networks of equations. A further advantage of these repertoire data (and equations) is that they can be searched and sorted very quickly in the database table.

Cluster Analysis: Cluster analysis sorts data into groups (clusters) according to stronger or weaker associations. It can be especially useful for comparing graphically similar sets of biological parts in different experimental settings. It extends our current ability from following the behavior of a few variables to one of looking at the larger patterns being produced by many variables.

Cluster analysis is a multivariate technique that provides graphical output as trees (dendrograms). Parts on the same branch are more closely related than those on distant branches. The results given in the software were calculated with StatistiXL (Version 1.6). An Internet search will provide enough background and examples to introduce the method to the reader (see, for example, the Clustan website).

When looking for mathematical patterns in the large data sets coming from microarrays, molecular biologists often use cluster analysis. A paper by Eisen et.al., (1998) includes an illustration and a worked example.

Hybrid Hierarchy Equation

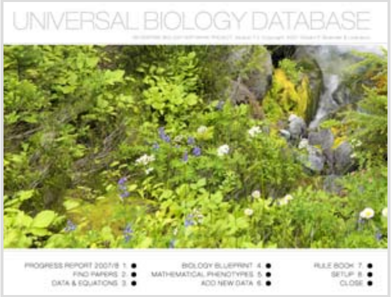
The hybrid hierarchy equation connects the first fourteen levels of the biological hierarch with the remaining two, namely molecules and genes. Making this connection, however, requires that the equation of the experiment uses both structural (stereology) and functional variables (biochemistry or molecular biology), and that all the data are collected as part of the same experiment.

This equation makes two noteworthy contributions. It allows the data of biochemistry and molecular biology to be upgraded to a quantitative level by providing valid biological interpretations for *in vivo* experiments and points out the importance of gold standards when developing and testing the future generation of structure-function methods. A brief introduction to this new equation can be found in he discussion and in chapter 6 of the **Rule Book** (Bolender, 2007).

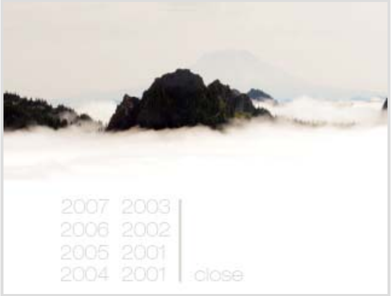
Methods and Results

Enterprise Biology Software Package – 2007

The software package includes a main menu sitting atop eight program modules. Each module provides access to a collection of programs, along with an introductory document (Read). The installation program installs the relational database, front-end tools, and includes supporting documents and programs. To run a program module, click on an item in the menu. Note that the text documents will display only after the new reader has been installed (see Setup 8.).   Since several of the programs were described in detail last year (Bolender, 2006), that effort will not be duplicated here. For more information about a given program, click on the **Read** button.



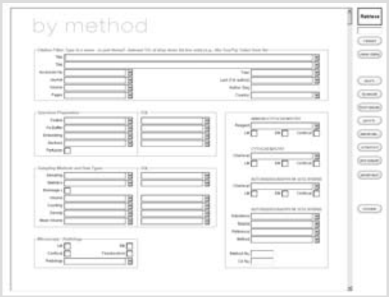
1. Read Progress Reports 2007/2008: Current and past reports can be called from this screen.



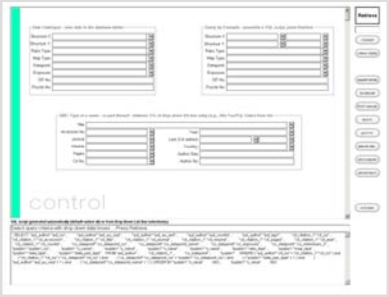
2. Find Research Papers: References can be found by searching on methods and on data. The **read** document provides examples.

By Citation:

By Method:

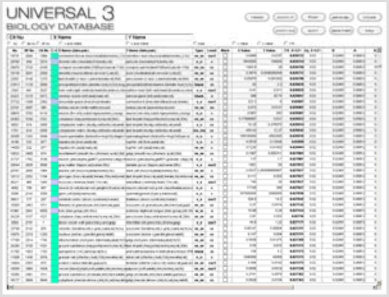


By Data:

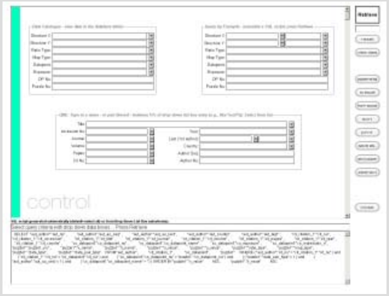


3. Explore with Data & Equations: The ***Universal Biology Database*** includes roughly 40,000 data pairs fitted to regression curves and summarized as decimal repertoire equations. This year, new data were added from about 100 papers. The module offers a variety of options for using these data in a discovery mode. The programs use a SQL (structured query language) front-end, one that can be mastered quickly even by beginners.

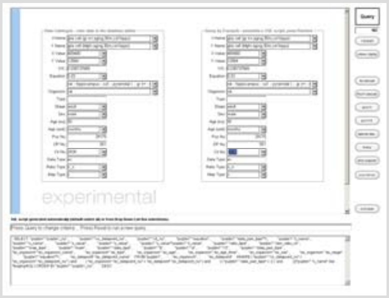
Data table (UBD):



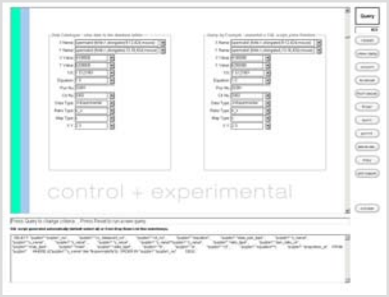
SQL – control data:



SQL – experimental data:

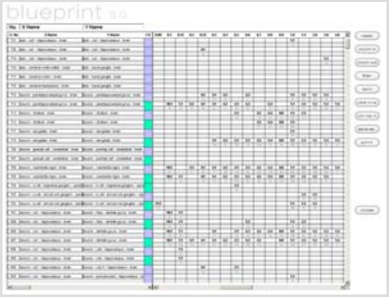


SQL – control & experimental data:

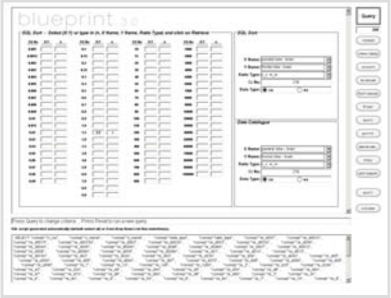


4. Find Quantitative Patterns with a Biology Blueprint: The blueprint summarizes the 40,000 decimal repertoire equations as a connection matrix. It shows how biological parts are connected quantitatively and how these parts can change in an experimental setting. In effect, it uses the proportion of parts (stoichiometry) to summarize the range of design options being employed by biology. This year, the biology blueprint was updated to include experimental data.

Blueprint out: The new version (3.0) includes both control (green) and experimental data (blue).

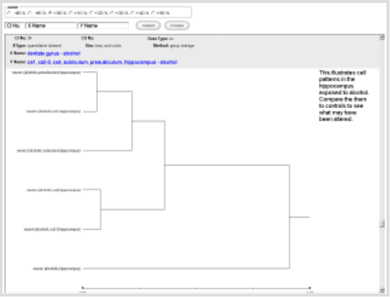


Blueprint SQL: A new SQL interface quickly locates specific information and patterns. Such tools should simplify the task of assembling networks and may become helpful in deciphering genetic switching and control mechanisms.



5. Viewing Mathematical Phenotypes: Decimal repertoire equations and cluster analysis sort data pairs into groups (or clusters) according to strong or weak associations. They offer effective ways of finding patterns in large and small data sets.

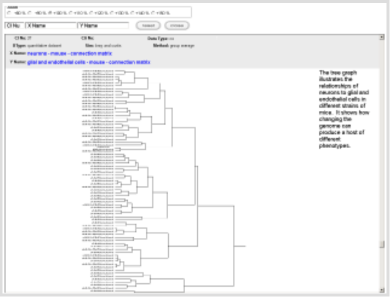
Comparing patterns: This module includes examples of tree graphs (dendrograms) and suggests a wide range of applications for the method. The figure at the left (below) illustrates the phenotype of the hippocampus of alcoholics, whereas the one at the right shows the different strains of mice used earlier in the connection matrix (Bolender, 2005).



X Name: dentate gyrus - alcohol

Y Name: ca1, ca2-3, ca4, subiculum, presubiculum, hippocampus - alcohol

This illustrates cell patterns in the hippocampus exposed to alcohol. Compare the them to controls to see what may have been altered.

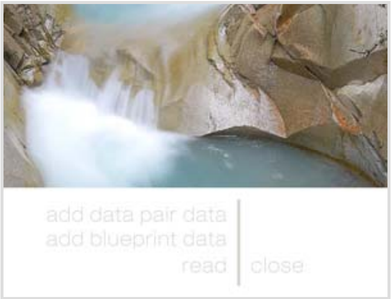


X Name: neurons - mouse - connection matrix

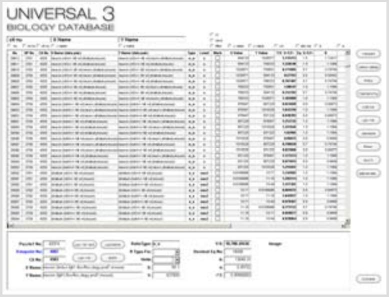
Y Name: glial and endothelial cells - mouse - connection matrix

The tree graph illustrates the relationships of neurons to glial and endothelial cells in different strains of mice. It shows how changing the genome can produce a host of different phenotypes.

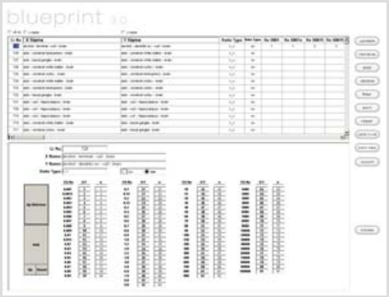
6. Add New Data: These data entry screens allow the user to enter new data pairs and blueprint data.



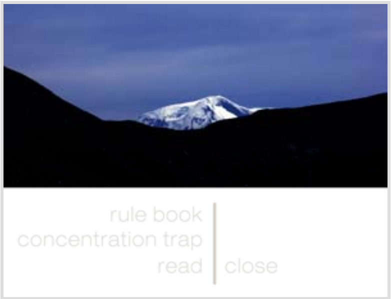
Data point data - in: Use this screen to enter new data pairs.



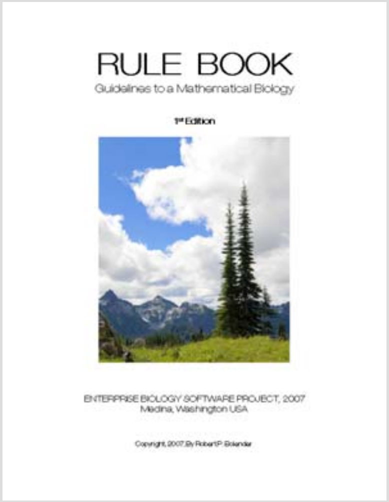
Blueprint in: Use this screen to enter new data into the blueprint.



7. Rule Book: The **Rule Book** introduces the ground rules for a mathematical biology based largely on the data and principles harvested from the literature of biological stereology. It also includes a software tool – called the concentration trap - that offers assistance in assessing the risk associated with using solo concentrations for detecting biological changes.



Rule Book: Guidelines for a mathematical biology.



RULE BOOK

Guidelines to a Mathematical Biology

1st Edition

ENTERPRISE BIOLOGY SOFTWARE PROJECT, 2007
Medina, Washington USA

Copyright 2007 By Robert P. Bolender

Concentration trap: What happens to the results of an experiment when it is based on an equation or on a single variable plucked out of an equation? For example, an outcome labeled V, S, L, and N all come from an equation, whereas Vv, Sv, Lv, And Nv represent plucked variables. Recall that the color-coding in this table includes: red an increase, blue a decrease, and green no change.

A global view: For example, click on the global button marked V**v vs. V** to see how often an isolated variable (Vv) – a concentration - gives the same result as that of the equation to which it belongs (V). Scroll through the screens - by clicking on them with the mouse - and become convinced that V does not always share the same background color with Vv. This means that an isolated variable cannot be trusted to give the same result as that of a properly written equation. On average, the two different outcomes agree only about 50% of the time.

A local view: All cells and tissues no not necessarily share the same risk of falling into the concentration trap. For example, counting the number of glomeruli in the kidney (Nv vs. N) is accompanied by 86% error (the concentration data will be wrong 86% of the time (24 disagree/28 total count), whereas counting neurons in the hippocampus (Nv vs. N) has only a 22% error (18 disagree/82 total). In any case, comparing densities or concentrations must be considered a risky business and all such comparisons should be considered highly questionable. Why? Because a single isolated variable can never be as effective at detecting a change reliably as a group of variables working together in a properly designed equation.

Here is the filter script for the glomerulus (match(ex_datapoint_ex_datapoint_name, '[g][l][o][m][e][r]') and ex_data_1_ex_n >0 and ex_data_1_ex_nv >0) and the one for the hippocampus (match(ex_datapoint_ex_datapoint_name, '[h][i][p][p][o]') and ex_data_1_ex_n >0 and ex_data_1_ex_nv >0). Direction for writing such scripts can be found in the software (click on Read).

8. Setup: Use the module to install the latest version of the Adobe reader on your computer and to set the path for the Excel files.

Discussion

One of the central challenges of modern biology and of this project is to understand how mathematics and technology can allow us to manage complexity in biology – locally and globally. One thing now appears fundamental to this process. Managing complexity consists of identifying the right collection of pieces that must be in place before a solution can emerge. Moreover, the difficulty of a given undertaking becomes proportional to the number of pieces involved. For example, figuring out how to design and populate the **Stereology Literature Database** was a singularly frustrating task because the database design worked only after it had been carefully aligned to the existing biology literature. In other words, the final database design ultimately came from the biology by way of the literature.

Similarly, many pieces must be in place before a mathematical biology can emerge from the biology literature. Once again, the data of the literature drives a solution to the problem of understanding the intrinsic order of biology. In this case, however, many of the pieces are far more intricate in design and subtler in action. The **Rule Book** catalogues these pieces, describing what they do and why they are important. The current collection of rules begins to trace the boundaries and capabilities of a mathematical biology that can be seen by peering at biology through a stereological lens (Bolender, 2007).

Semiquantitative Data

Today, the major argument for using semiquantitative data in the life sciences is the absence of quantitative alternatives. In turn, this argument underscores the absence of a theory structure upon which a mathematical biology can be built (National Academy of Sciences, 2008). The stereology community, however, has begun the process of designing and implementing an advanced research biology based on quantitative foundations. By applying mathematics and technology to the literature of biological stereology, the **Enterprise Biology Software Project** highlights and extends these accomplishments. Indeed, semiquantitative data today are no longer a practical necessity, but rather a second-rate option – albeit one not to be warmly recommended.

Of one thing we can be certain. Research biology needs a collection of rules or standards on the basis of which we can make informed decisions. Rules define the way we play the game – the way we set up and solve our problems in research. Today, even a cursory reading of the biology literature uncovers a wealth of semiquantitative data openly acknowledged or merely masquerading as quantitative data. Semiquantitative data do not play by the rules and invariably give confusing and conflicting results. Since biochemistry and molecular biology have become avid consumers of semiquantitative methods, our task of incorporating the data of these disciplines into the quantitative framework designed around stereology becomes one of figuring out how these data can be standardized and upgraded to meet the requirements of the database for data entry.

Why are semiquantitative data so popular? They continue to survive with so little effort because they generate numbers that can be shown to be significantly different and their flaws can be so easily dismissed. “Follow all those variables of immediate interest while holding constant (or ignoring) the remaining variables” makes remarkably good sense to someone faced with the task of exploring a system of enormous, yet largely unknown complexity. Indeed, the great power and appeal of semiquantitative data comes from the fact that they are largely an unknown quantity. As such, they promote a comfortable feeling in that one is under no obligation to explain the results of an experiment ***rigorously*** when it relies importantly on semiquantitative data.

If, however, we extend accountability to semiquantitative data, then they begin to loose some of their luster. How can we do this? It’s really quite straightforward. Since we now know from stereology how to translate experiments in biology into equations, we can employ a two-step process. First, we translate a **semiquantitative experiment** with its “variables of interest” into one or more equations and then expand these partial equations into a fully **quantitative experiment** supported by equations containing ***all*** the required variables. When we feed our experimental data to the two different sets of equations, we can learn - first hand - the difference between a quantitative and semiquantitative approach. What is the point? The point is a question. Why do we continue to do experiments that produce semiquantitative data when we no longer have to?

What are some of the experimental conditions that lead to semiquantitative data?

- Data are not collected with unbiased sampling methods.
- Data are collected with model-based methods.
- Experiments are based on partial or unbalanced equations.
- Data are expressed solely as densities or concentrations.
- Data are not collected and interpreted in ways consistent with the principles of biology.
- Data are collected in one dimension and interpreted in another – in the absence of a mathematical transformation.

Inconsistent Data Types

One of the basic understandings to come from this project is that we are creating non-natural phenotypes (e.g., *in vitro* conditions, transgenic animals) that display very different properties – locally and globally. This means that such artificial phenotypes appear inconsistent with the larger body of research data coming from natural sources. Mixing data from *in vivo*, *in vitro*, and transgenic studies can be expected to add considerable noise and thereby limiting our ability to find patterns. Indeed, it may be prudent to treat *in vivo*, *in vitro*, and transgenic experiments as distinct phenotypic categories and to characterize them separately as mathematical phenotypes.

Hybrid Hierarchy Equations

The richness of biology appears so great that it cannot be captured by a single discipline, no matter how powerful or successful it might be. Instead, the development of a mathematical biology requires a community effort – one being defined by all those disciplines capable of contributing variables to the equation(s) of an experiment.

It appears that solving a problem in research biology is not unlike building a winning team. Define the positions and then recruit the top players. Biochemistry and molecular biology can do an excellent job at counting the total number of molecules in a structure, but they are ill equipped to interpret these data in a complex *in vivo* setting and have little experience with design-based methods. On the other hand, stereology struggles when trying to count molecules but excels at dealing with complexity and in applying design-based methods. Put these methods together and we get our winning team. In short, the ***hybrid hierarchy equation*** offers a general solution to several problems. It can serve as the equation of an experiment, become a development platform, or assume the role as a gold standard.

The hybrid hierarchy equation derives its power from a curious property. In contrast to a regular hierarchy equation that is evaluated to provide an answer, the hybrid equation doesn't have to be evaluated because all the primary variables are collected experimentally – within the framework of a design-based model. The singular advantage of this approach is that it allows us to avoid many of the limitations and assumptions often accompanying individual methods.

Let's look at an example. The hybrid hierarchy equation for counting molecules with biochemistry and stereology in a biological setting is a follows.

$$N_{\text{molecules,structure}} = V_{\text{structure}} \cdot ((\text{mean}V_{\text{cell}} \cdot N_{\text{cell}}) / V_{\text{structure}}) \cdot N_{\text{molecules}}/V_{\text{cell}}$$

This equation can be used to detect an *in vivo* change in the number of molecules and to explain the change within a hierarchical setting. However, we can rearrange the equation and solve for the molecular concentration – shown as a numerical density ($N_{\text{molecules}}/V_{\text{cell}}$).

$$N_{\text{molecules}}/V_{\text{cell}} = N_{\text{molecules,structure}} / \{V_{\text{structure}} \cdot (\text{mean}V_{\text{cell}} \cdot (N_{\text{cell)}} / V_{\text{structure)}}))\}$$

Recall that concentrations are the principal type of data being collected by our experimental methods, especially those of biochemistry and molecular biology. Observe, however, that the concentration of this equation ($N_{\text{molecules}}/V_{\text{cell)}}$ was not measured directly but instead derived from two independent sources – biochemistry and stereology. In effect, this unique numerical density (concentration) becomes the gold standard for our numerical density estimates being routinely collected in the lab with many different methods and machines. In other words, it identifies a new strategy for developing, checking, and tuning any biochemical or immunocytochemical method for counting molecules that relies on data collected as a concentration (e.g., an optical density). See Chapter 6 in the **Rule Book** (Bolender, 2007) for further details.

Reverse Engineering – Keeping Score

Last year you may remember that biochemistry and molecular biology did not appear to be viable candidates for reverse engineering biology because the assay methods of these disciplines forfeit most of the structural information (Table 1). Recall that the task of reverse engineering biology is largely a structural exercise. Out of a possible twenty-eight dots, they got only six.

Table 1. Report 2006: Minimum requirements for reverse engineering biology using published data; a preliminary assessment. Can we fill in the missing dots? (Adapted from 2006 report.)

Requirements for Reverse Engineering	Stereology	Biochemistry	Molecular Biology
In Vivo Data			
Concentration Data	●	●	●
Average Cell Data	●		
Absolute Data	●	●	
Cell Counts	●		
Molecule Counts	●	●	●
Minimize Bias	●		
Minimize Animal Variability	●		
Detect Change Unambiguously	●	●	

Design Experiments as Equations	●		
Enforce Unbiased Sampling	●		
Apply Biological Rules	●		
Convert 2D Data back to 3D	●		
Standardize Data	●		
Generate Biological Blueprints	●		

As you can see in the scorecard this year (Table 2), biochemistry and molecular biology look a good deal more promising in that the blue dots identify the likely outcomes when these disciplines join forces with stereology within the framework of the hybrid hierarchy equations.

Table 2. Report 2007: Minimum requirements for reverse engineering biology using published data; a preliminary assessment. Can we fill in the missing dots? (Adapted from 2006 report and updated.)

		Stereology +	Stereology +
Requirements for Reverse Engineering	Stereology	Biochemistry	Molecular Biology
In Vivo Data			
Concentration Data	●	●	●
Average Cell Data	●	🔵	🔵
Absolute Data	●	●	🔵
Cell Counts	●	🔵	🔵
Molecule Counts	●	●	●
Minimize Bias	●	🔵	🔵
Minimize Animal Variability	●	🔵	🔵
Detect Change Unambiguously	●	●	🔵
Design Experiments as Equations	●	🔵	🔵
Enforce Unbiased Sampling	●	🔵	🔵
Apply Biological Rules	●	🔵	🔵
Convert 2D Data back to 3D	●	🔵	🔵
Standardize Data	●	🔵	🔵
Generate Biological Blueprints	●	🔵	🔵

Chaos Theory – A short Story

Let’s change the subject. A story widely circulated about chaos theory is the one about a butterfly that could change the weather patterns throughout the world by simply flapping its wings – known as the “butterfly effect.” Of course, such an event seems quite unlikely, but it nonetheless raises an intriguing question. What might we do for this butterfly to improve its chances of success in changing our global weather patterns? Think a moment. If we draw an analogy to chaos theory, then all we might have to do is move the butterfly to the edge of chaos where such events routinely occur as emergent properties. How can we do this? I don’t know. However, the question provokes another one to which we might hazard an answer.

Consider this. Is reverse engineering biology an emergent property of the biology literature? If the answer is yes, then how might we move our research data to the edge of chaos to activate this emergent property? Now we have a question to which we might have an answer. One way seems to consist of moving our published data into a **Universal Biology Database** and then using it to generate the engineering equations. Look back at the progress reports to see what has already emerged. This universal database allowed us to summarize a large collection of control and experimental data with only two equations, to use published data to forward and reverse structures, and to display the mathematical core of biology as a stoichiometry blueprint. Does this mean that a sizable portion of the stereology literature is now sitting at the edge of chaos? I don't know. But if it is, then it might be exactly at the "sweet spot" where a host of new and quite unexpected things are about to happen.

Concluding Comments

Where does biology go to discover things? It goes to the edge of chaos. Why? Because that's where emergent properties occur. These properties are the discoveries. Can we copy this discovery strategy of biology? Perhaps we can because stereology seems to be unusually clever at opening tightly closed doors - mathematically. How? One strategy might be to move more of the published data of research biology – from many different disciplines - to the edge of chaos as well. How do we do that? Use more mathematics and technology. How will we know when we have made it to the edge? We will begin to discover curious things for the first time and know exactly what to do next. In other words, we will begin to experience what it means to be wired into the mathematical core of biology.

References

Board on Life Sciences, National Academy of Sciences. 2008 The role of Theory in Advancing 21st Century Biology: Catalyzing Transformative Research, Washington D.C.: National Academies Press.

Bolender, R. P. 2001a Enterprise Biology Software I. Research (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 © 2003 Robert P. Bolender

Bolender, R. P. 2004 Enterprise Biology Software V. Research (2004) In: Enterprise Biology Software, Version 4.0 © 2004 Robert P. Bolender

Bolender, R. P. 2005 Enterprise Biology Software VI. Research (2005) In: Enterprise Biology Software, Version 5.0 © 2005 Robert P. Bolender

Bolender, R. P. 2006 Enterprise Biology Software VII. Research (2006) In: Enterprise Biology Software, Version 6.0 © 2006 Robert P. Bolender

Bolender, R. P. 2007 Rule Book: Guidelines to a Mathematical Biology (2007) In: Enterprise Biology Software, Version 7.0 © 2007 Robert P. Bolender

De Duve, C. 1974 Nobel Lecture: Exploring cells with a centrifuge. From Nobel Lectures, Physiology or Medicine 1971-1980, Editor Jan Lindsten, World Publishing Co., Singapore, 1992.

Eisen, M. B., Spellman, P. T., Brown, P. O., and D. Botstein. 1998 Cluster analysis and display of genome-wide expression patterns. PNAS Genetics 95: 14863-14868.