Enterprise Biology Software: X. Systems Biology Two (2009)

ROBERT P. BOLENDER

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

Summary

By asking where it comes from rather than what it is, we soon discover how **Systems Biology Two** fits into the larger scheme of **Enterprise Biology**. This question drives the next chapter of our story. The report shows that if we begin by optimizing the effectiveness of our methods and data, then we end up with an optimized information infrastructure for biology. In turn, this infrastructure - defined explicitly by data, software, and outcomes - becomes a springboard to systems biology. Notice the basic fundamentals of the infrastructure: unbiased data, digitized literature (biological stereology), standardized data, hierarchical organization, experiments as equations, gold standards, biological changes detected unambiguously, reproducibility, engineering (forward and reverse), data catalogues, universal biology database, scalability, data integration, biology blueprints, and quantitative phenotypes (connection and percentage change). By including all biological disciplines and most data types, we now have a broadly based platform for making discoveries and finding generalizations - both locally and globally. The report introduces this information infrastructure as an organizational chart, translates the chart into software, and then uses the software to explore models for systems biology. The first such model defines systems biology as a collection of quantitative phenotypes undergoing complex changes. These changes, which include steady state and transitional, alter both the amounts and proportions of parts as biology redefines itself. The reader quickly discovers the effectiveness of this approach in that it straight away increases our ability to detect biological changes routinely by amounts approaching an order of magnitude. Notice how this new discovery process works. We begin by watching biology change and then imitate it by applying technology. By also allowing the biology literature to redefine itself in parallel with biology, it too becomes highly productive, generating a host of reality-based phenotypes for development, disease, and experimental settings. Spreadsheet files offer simple, worked examples of how to transform published data into quantitative phenotypes displaying the complexity of biological changes. As part of the yearly report, contributing authors will receive copies of these files in their EBS software package. Notice, if you will, how the project approaches the overarching problem of biological complexity. Strategically, it considers the big picture, sets priorities, recognizes patterns, anticipates issues, predicts outcomes, and operates at both local and global levels. Tactically, it gets the job done by meeting the strategic goals, optimizing outcomes, applying best practices, and delivering new software tools promptly to the community.

Introduction

What exactly is **Systems Biology Two**? It represents an attempt to fill the gap created by systems biology, as it exists today. An Internet search on <systems biology definition> suggests that it focuses almost exclusively on genes, molecules, and networks, including a yet to be defined connection to individualized health care. The model looks like this: genes->molecules-> ,…, ->organism. If we call this model **Systems Biology One (SB1)**, then the gap represented by the three dots (,…,) becomes **Systems Biology Two (SB2)**. In combination, they begin to define a single, all-inclusive, systems biology.

Our story, however, begins not with systems biology, but with an information infrastructure. Such an infrastructure represents the application of technology to the biology literature wherein published data become allowed to interact freely and produce new forms of information. The first lesson to come from this new technology is that biology obeys rules that we can capture empirically by fitting published data to equations. In effect, technology allows us to assemble a standardized, integrated, and quantitative biology – based entirely on the literature. Such an outcome encourages the formation of an information infrastructure capable of connecting and moving data freely throughout the biological hierarchy of size – extending all the way from genes to organisms. This is exactly what one would expect from something called systems biology. To assure its success, however, the infrastructure must be very good at detecting biological changes accurately even when the complexity of these changes approaches the extreme. If systems biology is to become the business of detecting broken or defective parts, designing replacements or treatments, and evaluating the results, then it will no doubt benefit significantly from the many features being offered by this new resource.

The report introduces the reader to the infrastructure, beginning with its design and continuing with the databases, software tools, and results. Throughout the document, notice how carefully engineered features become essential to the effectiveness of an information infrastructure and the way in which information flows through the system, moving from data catalogues to discovery platforms. Biology takes center stage. Wherever we look quantitatively, biology delights us by displaying well-defined stoichiometries of its parts accompanied by a surprisingly fierce determination for maintaining healthy patterns. Three new products include two libraries for detecting change (steady state and transitional) and a systems biology library that catalogues changes produced by development, disease, and exposure. The principal challenges for the reader will be to work through the software and documents and to become familiar with the process of using technology to see how biology manages change to solve its problems. Our reward – as players – comes as an ability to move our game plan from simplicity to complexity – from reductionism to connectionism. Moreover, new opportunities arise for changing our perspectives on discovery. As a direct product of the information infrastructure, we can continually upgrade our model for systems biology by simply returning to its source and selecting existing features – already tested both locally and globally. Instead of continually reinventing the wheel, we simply help it to run faster and smoother.

Methods and Results

The software package for 2009/2010 includes new software tools for studying change (transitional and steady state) and begins the process of defining the capabilities of systems biology by generating local and global summaries of complex changes.

Enterprise Biology Software Package for 2009/2010

The package includes eight screens offering access to programs, databases, and documents. Taken together, they define an information infrastructure that will serve as our discovery platform for SB2.

Figure 1 Enterprise Biology Software Package for 2009/2010

One of the most striking features of the package is that it allows the user to move published data readily from one database to another, thereby creating new data types and applications. Indeed, the mainstay of any information infrastructure exists as an ability to extract new and useful information from old. The addition this year of more than 10,000 new database entries shows just how well this strategy works in practice.

Information Infrastructure

The organizational chart shown in Figure 2 begins with the **Biology Literature** and works its way down to the bottom box marked **Systems Biology Two**. The organizational plan rests on the assumptions that

we can collect high quality data and make interpretations consistent with reality, which includes detecting what actually happens biologically. In practice, the process of building out an information infrastructure consists of creating literature databases as data catalogues, which, in turn, can then serve as platforms for assembling the next generation of discovery databases. For example, the catalogue created by the Stereology Literature Database provided the data needed to assemble a Universal Biology Database, which, in turn, provided data for assembling the Digital Libraries. Upon reflection, one begins to understand how a thoughtfully designed and carefully tested information infrastructure can provide the solid foundation essential to the emergence of a broadly based systems biology.

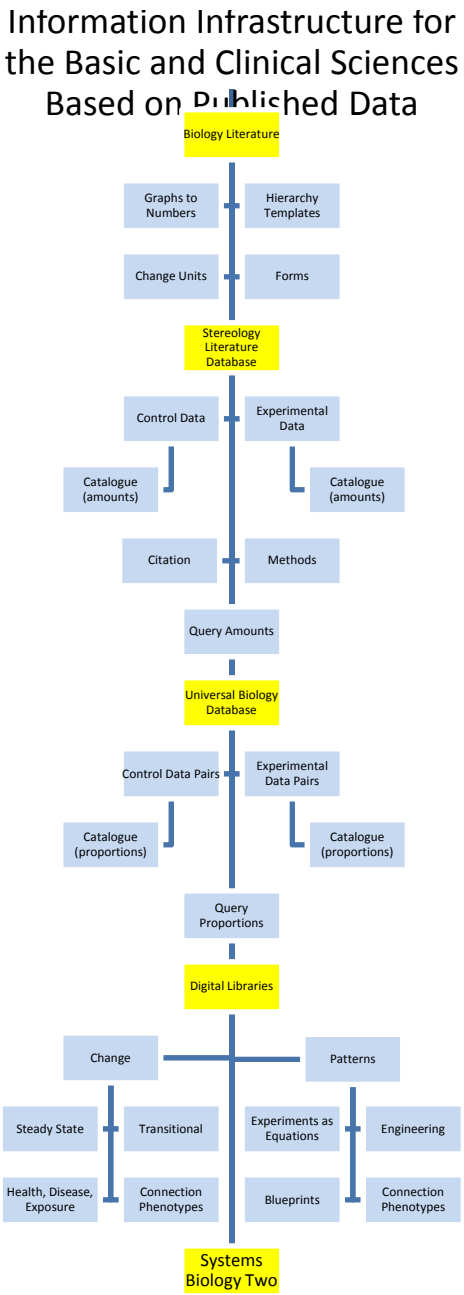


Figure 2 Organizational chart for the information infrastructure

For convenience, we can summarize the components of the organizational chart with a table of components and descriptions.

Table 1 Details of the organizational chart

Component	Description
Publications	Identifies refereed papers wherein data were collected with unbiased sampling methods and changes were detected unambiguously
Graphs to Numbers	Converts data reported graphically back into numbers
Hierarchy Templates	Defines a structural hierarchy for each data point to be entered into the database
Change Units	Provides a tool for changing data units
Forms	Aids for assembling data tables for data entry
Stereology Literature Database	Standardizes data, terminology, and hierarchical order; applies strict rules for data entry
Control Data Catalogue (Amounts)	Accommodates most forms of data from most biological disciplines
Experimental Data Catalogue (Amounts)	Accommodates most forms of data from most biological disciplines; identify significant differences
Citation	Includes complete references
Methods	Includes methods
Query Amounts	Uses a query-by-example (QBE) interface
Universal Biology Database	Minimizes biases and animal variability by forming data pairs; serves as the principal discovery database
Control Data Pairs Catalogue (Proportions)	Forms control data pairs for all possible combinations hierarchically, using data from the Stereology Literature Database
Experimental Data Pairs Catalogue (Proportions)	Forms experimental data pairs for all possible combinations hierarchically, using data from the Stereology Literature Database
Query Proportions	Uses a query-by-example (QBE) interface
Digital Libraries	Includes applications derived largely from the Universal Biology Database
Change – Steady State (related to control)	Change that may have reached a dynamic equilibrium; (1) data ratios (identified as DREs) plotted for control and experimentals; (2) each experimental data ratio expressed as a percentage of the control ratio
Change – Transitional (related to earlier time point)	Change that is in the process of changing; (1) data ratios (identified as DREs) plotted for controls and experimentals, wherein each data ratio is divided by the earlier - one starting with the control; (2) each data ratio expressed as a percentage of the earlier ratio
Change- Health, Disease, Exposure	Change captured by equations
Change-Connection Phenotypes	Change illustrated as complex curves
Patterns-Experiments as Equations	Experiments designed and interpreted as equations
Patterns-Engineering	Equations for reverse and forward engineering biological structures and functions
Patterns-Blueprints	Stoichiometry expressed as a frequency distribution of connections (data pairs) across biology
Patterns-Connection Phenotypes	Distinct curves for specific conditions (fingerprints)
Systems Biology Two	Everything connected mathematically

Information Infrastructure as a Software Tree

As shown in Figure 3 below, the process of building the information infrastructure consists of translating the boxes of the organizational chart of Figure 2 into a database, program, or document, and then arranging them conveniently on the branches of a software tree. To operate the information

infrastructure, the user simply opens the tree and selects those items of interest by clicking on the corresponding **GO** buttons. To display the name of the program selected, simply right click the heading.

Data entry initiates a remarkable journey. Once identified in a graph or table, each numerical value is activated by translating it into a digital form. As such, it can be standardized, organized, tidied up, and put to work. As it moves freely through the information infrastructure, it continually reinvents itself by joining equations or making new connections. Instead of contributing just a single piece of information, as determined by the immediate goals of an experiment, it enters into a continuous state of productivity. This tells us that our data, if given the chance, will continue to contribute to the formation of new knowledge for many years to come. In effect, data can become amaranthine.

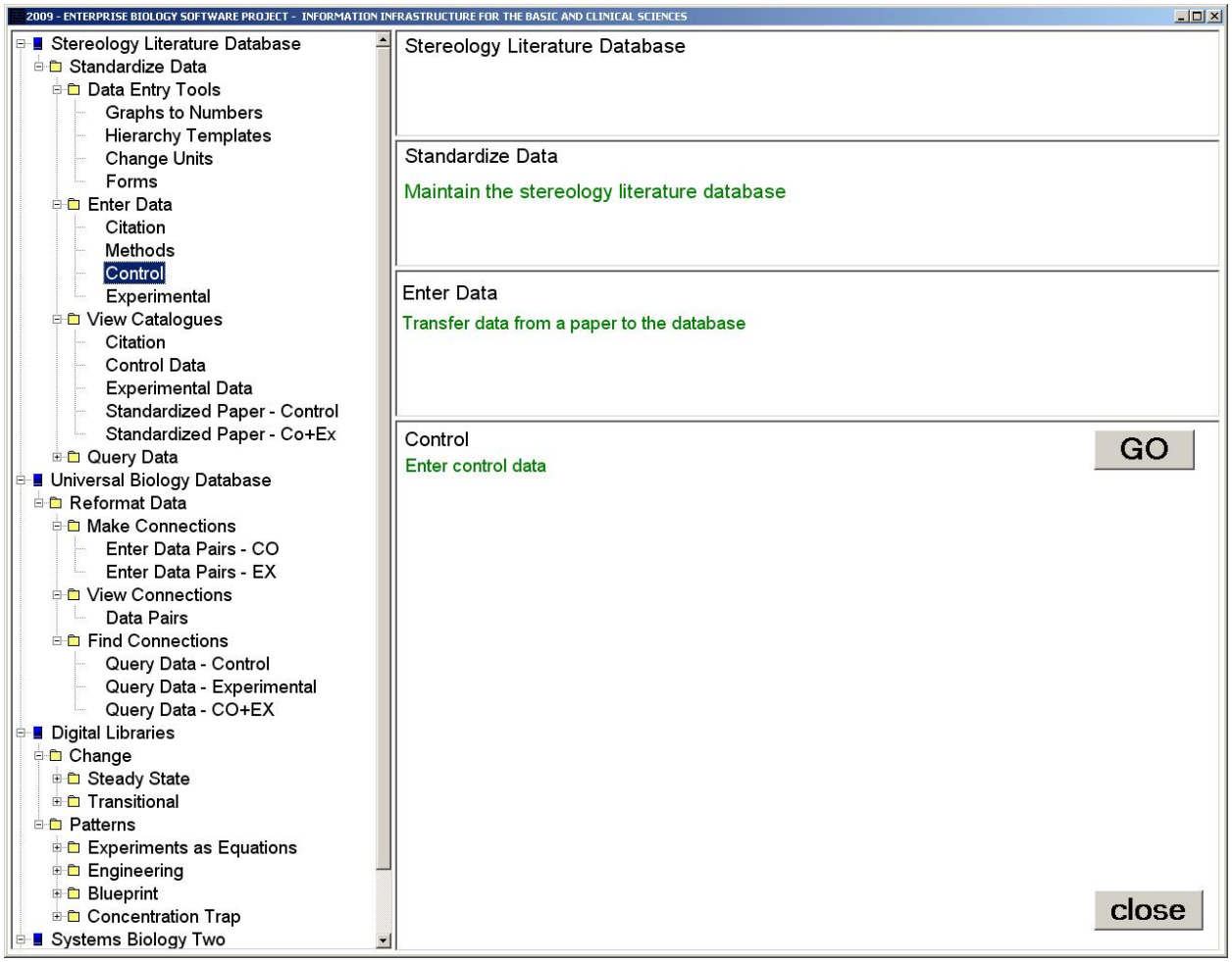


Figure 3 Software tree

The information infrastructure runs largely on equations, fed with data coming from unbiased sampling methods. Table 2 briefly summarizes several of these equations.

Table 2 Equations of the Connection Model

Equation	Description
Stereological Equations	Provide unbiased estimates for both amounts (volumes, surfaces, lengths, and numbers) and concentrations (volume densities, surface densities, length densities, and numerical densities) of parts
Hierarchy Equations	Connect data mathematically within and across hierarchical levels
Equations of the Experiment	Represent hierarchy equations wherein variables define the data to be collected and analyzed
Design Code Equations	Plots data from one time point to the next as a regression line, using only those point located on or adjacent to the line; identifies variables that change according to rule
Repertoire Equations	Data pairs (Y/X) with similar ratios fitted to a regression line having coefficients of determination (r^2) close to 1.0
Decimal Repertoire Equations	Repertoire equations grouped according to decimal steps and having coefficients of determination (r^2) close to 1.0
Ladder Equations	Summarize all the control and experimental data pairs as two intersecting exponential equations
Hybrid Hierarchy Equations	Combine structural and functional variables into a single equation of the experiment, thereby providing a direct experimental connection between the data of biochemistry, molecular biology, and stereology
Connection Phenotype Equations	Display collections of data pairs as a continuous curve (complex polynomial)
% Ratio Change Equations	Express two data pair ratios as a percentage; a steady state change compares an experimental time point to its control and a transitional change compares an earlier time point to a later one

Information Flow

One of the central challenges in creating an information infrastructure consists of allowing published data to flow freely throughout the system so that we can assemble them into new configurations. Such configurations become new forms of information. Figure 4 illustrates this process with three data entry screens, beginning with (1) original data entry, followed by two assembly steps: (2) forming data pairs and (3) mapping one data pair onto another.

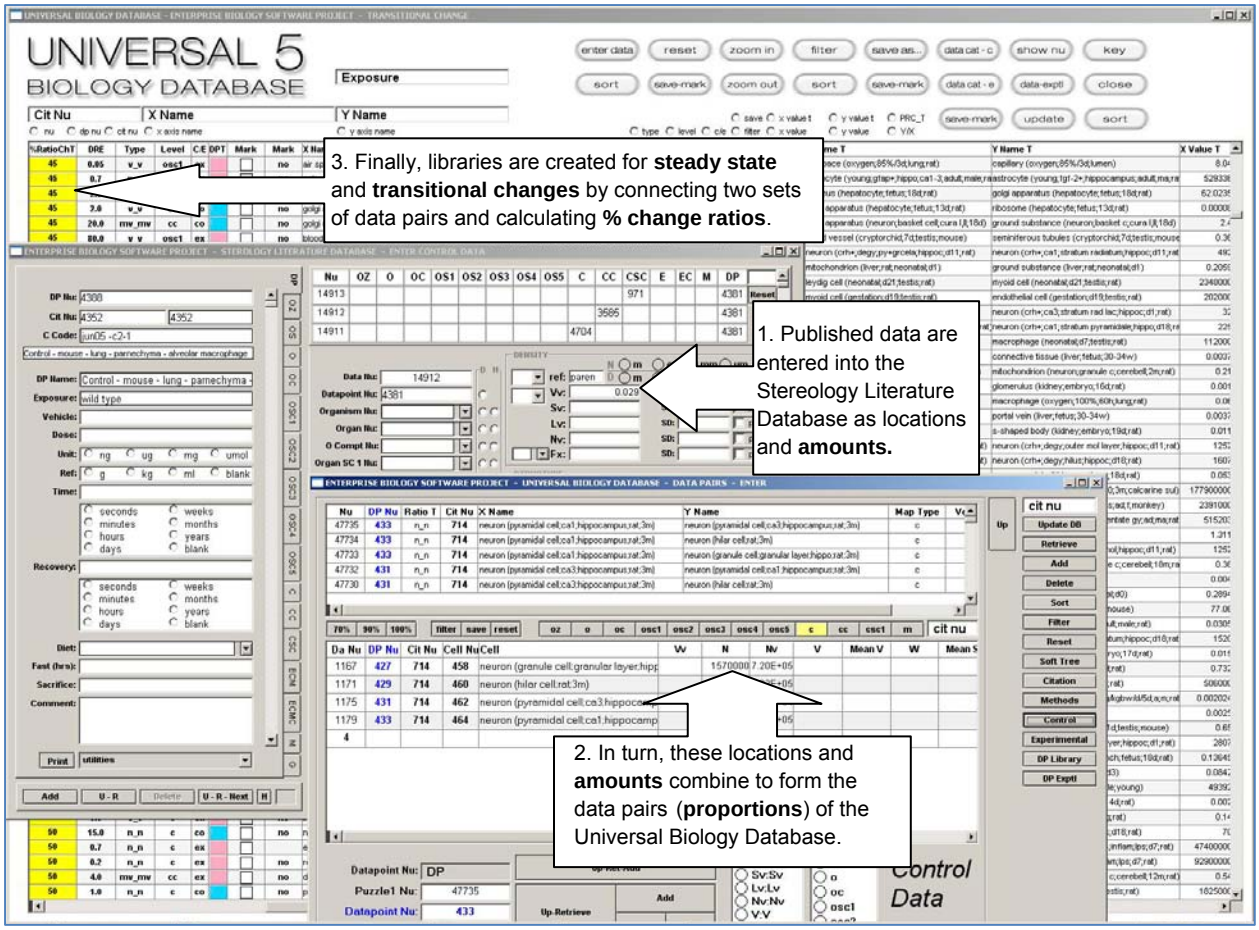


Figure 4 Information flow to and from relational databases

The process of establishing a free flow of information begins by collecting numerical data from a paper and arranging them in tables. To put these data back into the original context of biology, we need to assemble a structural hierarchy of locations for each part onto which we can map numerical values (amounts) – one such hierarchy accommodates controls the other experimentals. Since these hierarchies contains thirty-two levels of locations (16 control; 16 experimental) and at least thirty-two different types of numerical data, we rely on a relational database model to keep track of all the many parts and relationships. When translated into a user-interface, this database model becomes the data entry screen shown as item 1 in Figure 4.

Change

We – as biologists – tend to treat change as something that occurs at the level of biological parts or compartments, whereas biology orchestrates change by redefining its phenotype - locally and globally. Whereas we treat change as a statistical difference in the **amounts** of something, biology approaches

change by redefining both the **amounts and proportions** of its parts, using a rule-based system to optimize its phenotype. Notice the difference. We assume the luxury and convenience of picking and choosing what we do, whereas biology has to do all that is necessary to survive. This chasm - existing between convenience and reality - creates a troublesome dilemma often accompanied by unintended consequences. To Wit: we may be in the biology business, but not necessarily in the business of biology.

Changes in the amounts and proportions of parts define a major component of biological complexity. Therefore, an ability to mirror the many types of biological changes becomes a key function of the infrastructure and of systems biology. Table 3 lists and describes twelve methods currently available for detecting biological changes.

Table 3 Changes detected within the information infrastructure

Methods for Detecting Change	Description
Absolute Amounts	Volumes, surfaces, lengths, numbers, and weights of parts: V, S, L, N, W
Concentrations (Densities)	Amounts/unit of containing (reference) volume: optical density (N/V); V/V, S/V, L/V, N/V, Units/V, Units/W
Proportions (Data Pair (DP); Data Pair Ratio (DPR); DREs)	Ratios of absolute values: V:V, S:S, L:L, N:N, meanV/MeanV, meanS/MeanS, meanL/MeanL, meanN/MeanN
Design Codes	Plot sets of adjacent time points to get regression equations with $R^2 \approx 1/0$.
Connection Phenotypes	Display data pairs as frequency distributions
Connection Phenotypes: Based on concentrations or densities	Detect absolute changes in the number of data pairs
Decimal Repertoire Equations (DREs)	Identify the proportions of parts at different time points
Steady State Change	Compare experimental data pairs to corresponding controls
Transitional State Change	Compare the same data pairs of adjacent time points
% Change Ratios – Steady State	Compares a data pair ratio of the control to an experimental data pair ratio – useful for identifying dynamic equilibriums and plateaus
% Change Ratios – Transitional State	Compares a data pair ratio (control or experimental) to the data pair ratio of the adjacent time point – useful for identifying transitional changes

Change – Transitional and Steady States

The organizational chart (Figure 2) shows that most of the quantitative power of the information infrastructure comes from the data pairs (proportions) of the Universal Biology Database. Data expressed as ratios (Y/X) allow us to recreate the local and global patterns associated with connectivity and change. In fact, the strategy being developed for **SB2** depends importantly on identifying changes in the individual ratios of phenotypes over time – to which we now direct our attention.

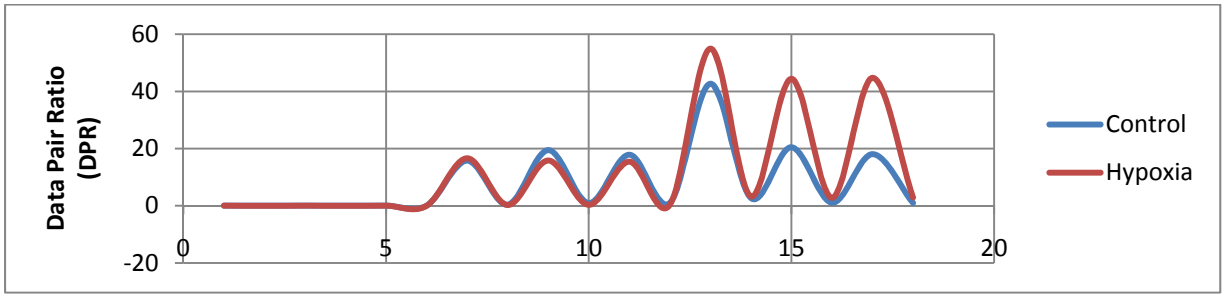
We can capture a biological change with proportions as it occurs (transitional state) or at its completion (steady state). Typically, a steady state change divides an experimental data pair ratio by its control ratio, whereas a transitional change divides the data pair ratios of adjacent time points. In practice, a steady state exists when an experimental curve returns to that of its control state, when two adjacent curves superimpose, or when data exist for only two time points (control and experimental).

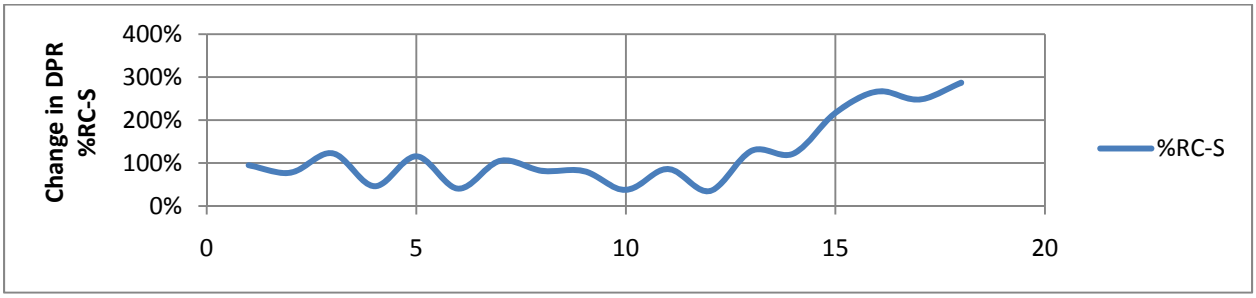
Transitional and steady state data also report changes in the data pair ratios as percentages. The process of assembling this tool consists of adding new columns (Name_X_i, Name_Y_j, Value_X_i, Value_Y_j, Value_$Y_{j/}$ Value_X_i) to the table of the Universal Biology Database. For the steady state modification, control data populate the new columns; whereas the data of previous data points (time point minus one) populate the transitional modification. Calculation fields supply the individual data pair ratios (DPR) and report a change therein as the % Ratio Change (%RC-T for transitional and %RC-S for steady state).

Both the transitional and steady state tables include two sets of data pair ratios (DPRs) that can be plotted individually as connection phenotypes and together as **% Ratio Changes**. These tables allow us to identify data pairs that change and to see by how much. Two points quickly emerge from this method of presenting data. Changes in proportions occur far more frequently than changes in amounts and often appear in places we rarely look.

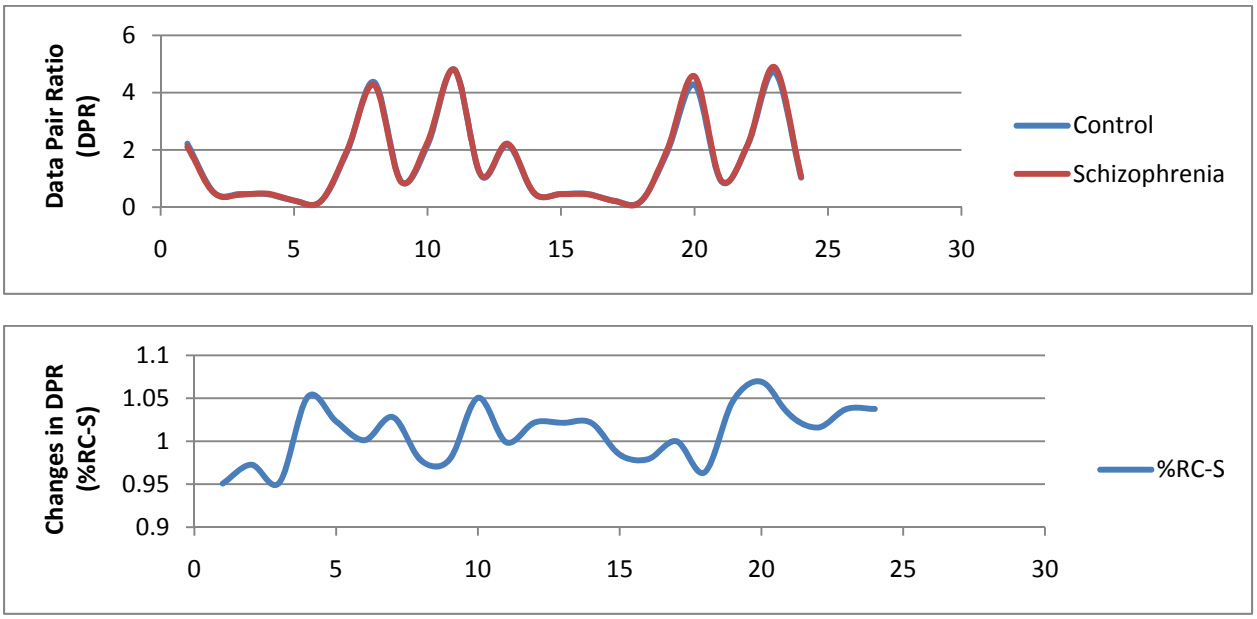
Worked examples - selected from the literature - will serve to illustrate the effectiveness of this approach to detecting changes in biological phenotypes.

Steady State Change: Response of brain parts to hypoxia (Cit 2312). Notice the progressive increase over time. (N.B., The horizontal axis plots the contents of a sorted data table by row number.)





Steady State Change: Schizophrenia of the human hippocampus (Cit 230). Notice the superimposition of the two curves. Although this suggests the absence of change, it also demonstrates the remarkable ability of the methods and workers to detect largely identical phenotypes in two distinctly different populations of patients. Notice also that the % Ratio Changes (%RC-S) indicate that the two estimates (control vs. schizophrenia) differ on average by less than five percent, a pattern often seen in many of the worked examples of the SB2 folders.



Steady State Change: Schizophrenia of the human hippocampus (Cit 3489). In this study, we also see similar patterns for both sets of patients. However, the change data (%RC-S) suggest widespread, but low level changes. This could be the result of the methods or it may be telling us that proportional data are more effective than absolute when studying these changes related to schizophrenia.

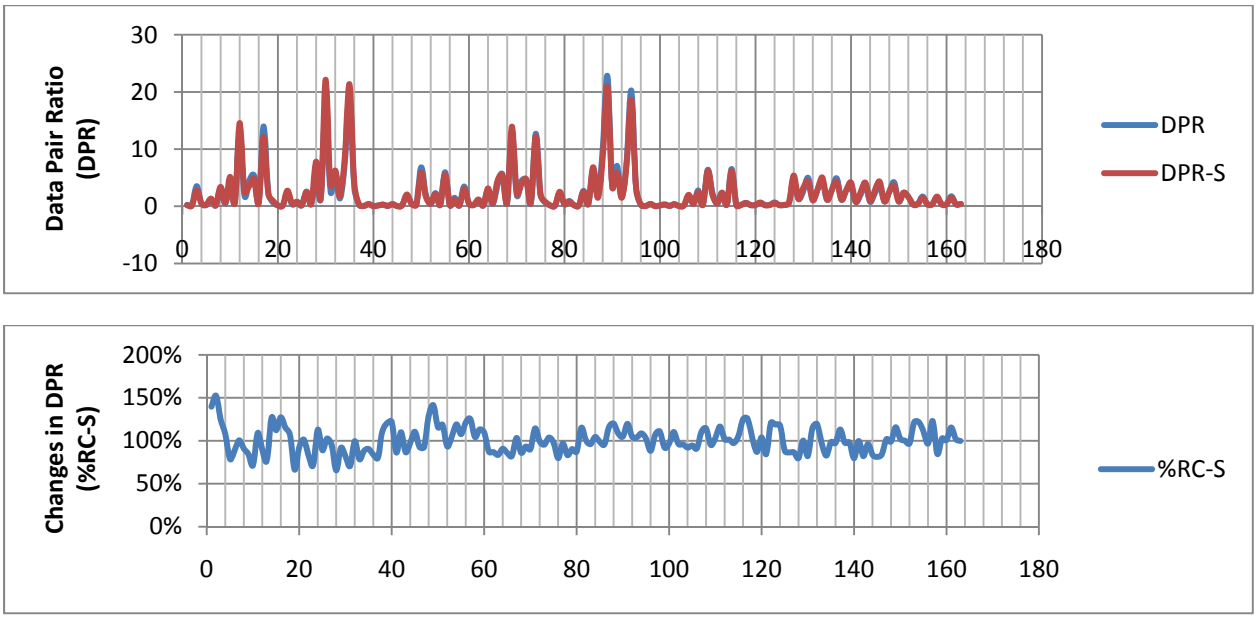


Figure 5 allows us to explore these two possibilities. In this study of schizophrenia, no significant differences were reported for the number of neurons in the human hippocampus. Notice that the <0.05 column next to N (numbers) of the Stereology Literature Database is blank. However, when we look at the proportions of these same neurons in the Universal Biology Database, many changes (49) suddenly appear. Values marked with a yellow background identify % Ratio Changes at ±15% (≥ 115% and ≤85%). This shows us that biology effectively remodels the brain without introducing absolute changes large enough to be detected by our traditional experimental methods. Proportional data are simply more sensitive to change than absolute data. Apparently, minimizing biases and animal variations would seem to move us closer to reality.

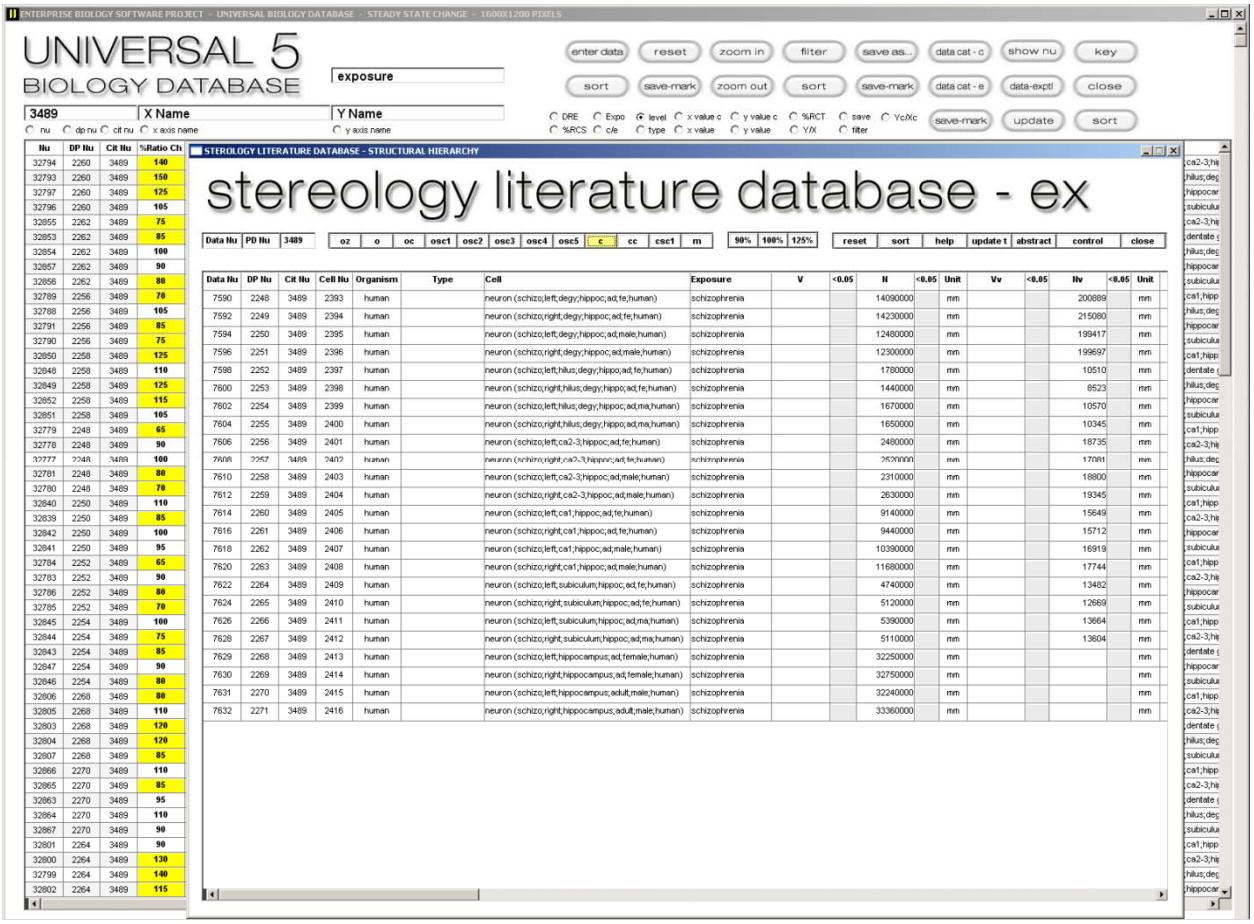
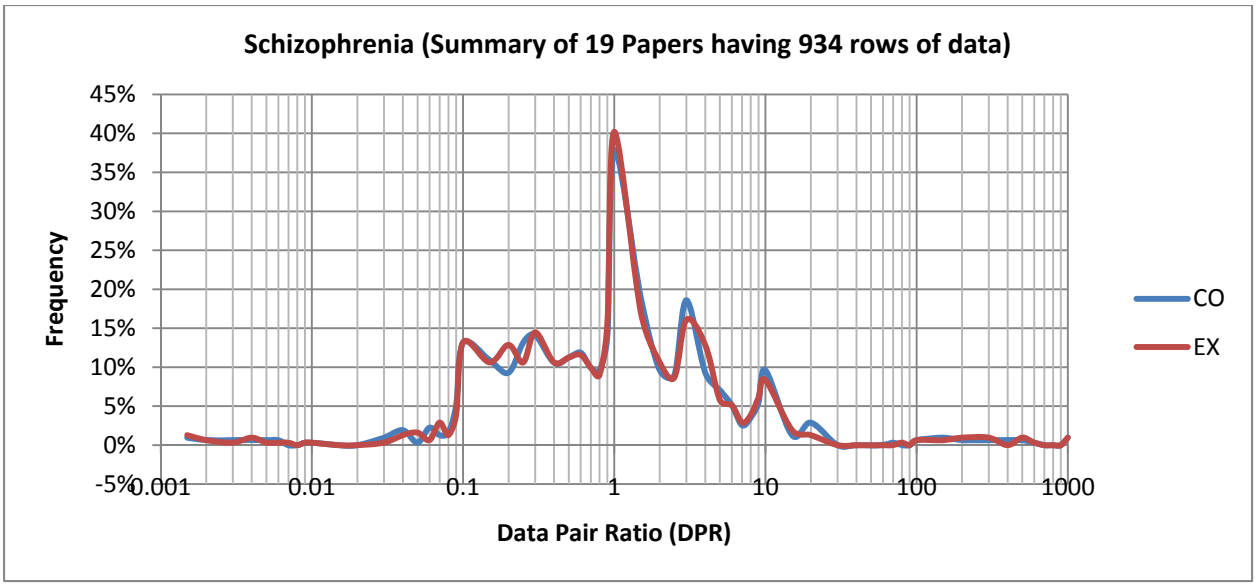
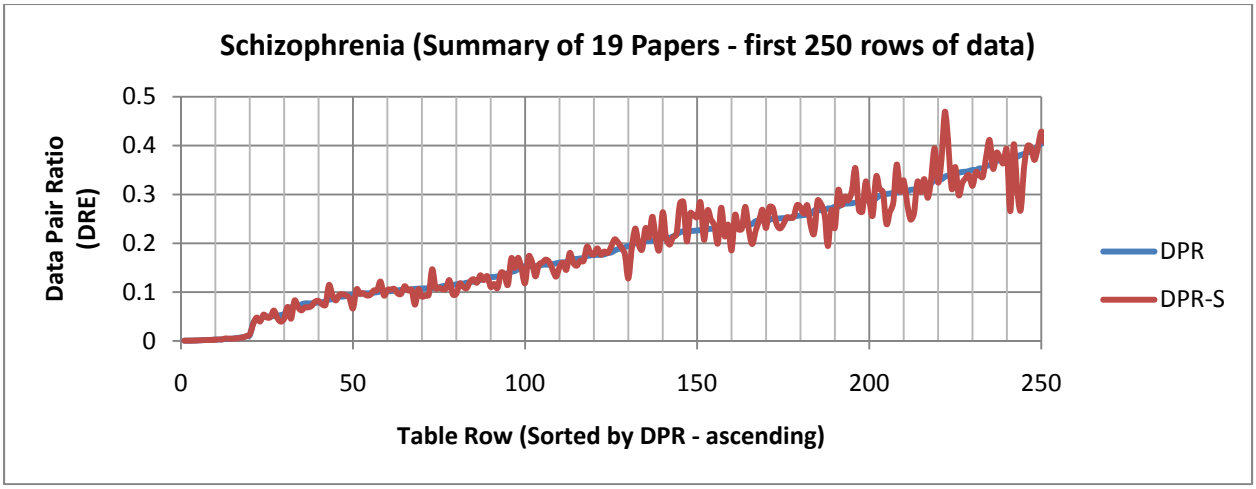


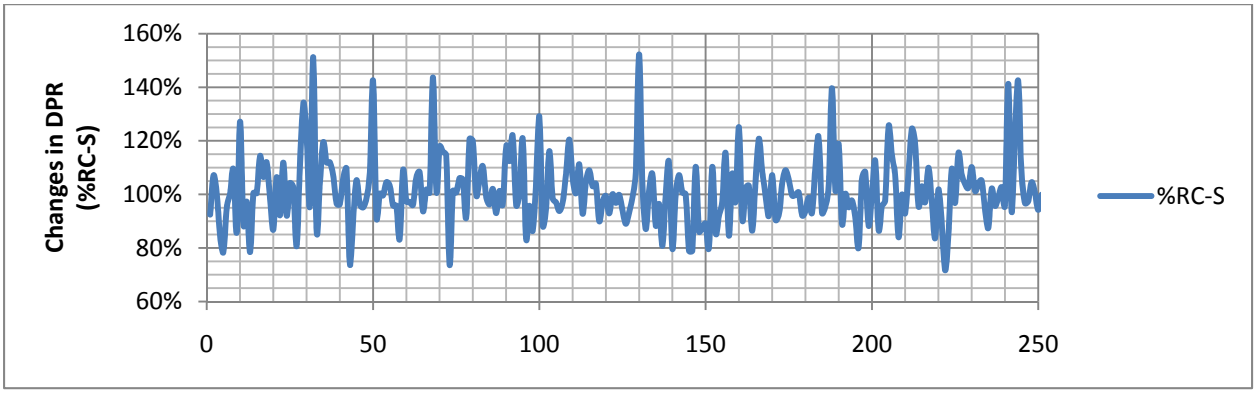
Figure 5 Detecting "hidden" changes with proportions.

While we study schizophrenia one or a few parts at a time, biology deals with it at the level of the brain and that of the entire organism. What happens when we ask, "How does schizophrenia affect the human brain?" The infrastructure allows us to answer such a question by first integrating the data of several research papers (19) and then displaying the results as a global connection phenotype. Now we can see changes occurring throughout the brain.

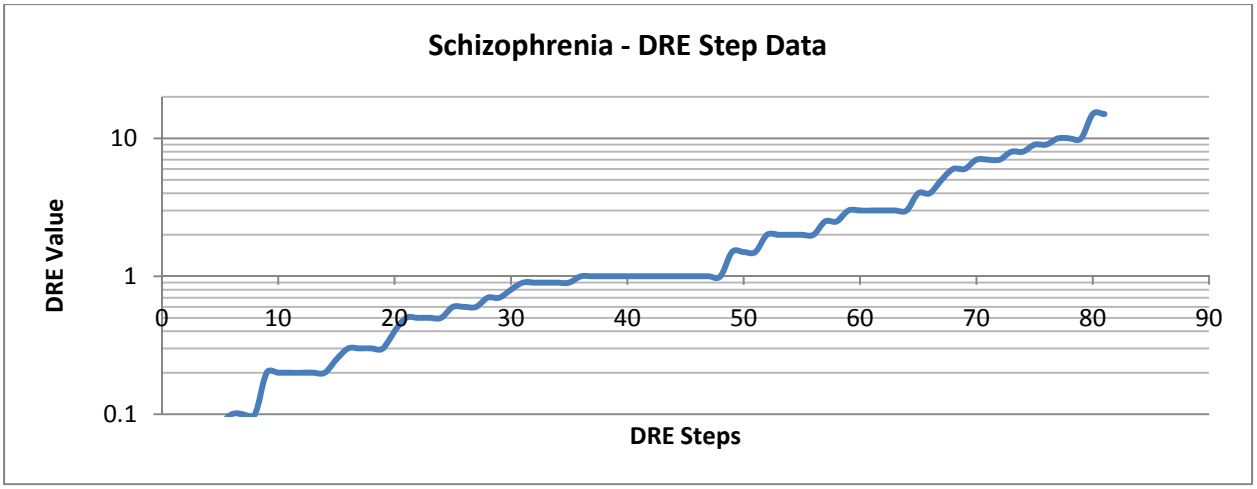


If we sort this data table by ascending data pair ratios (DPR) of the controls (blue line), then we can see how the data pairs changed in schizophrenia – as ratios and as percent ratio changes (%RC-S).

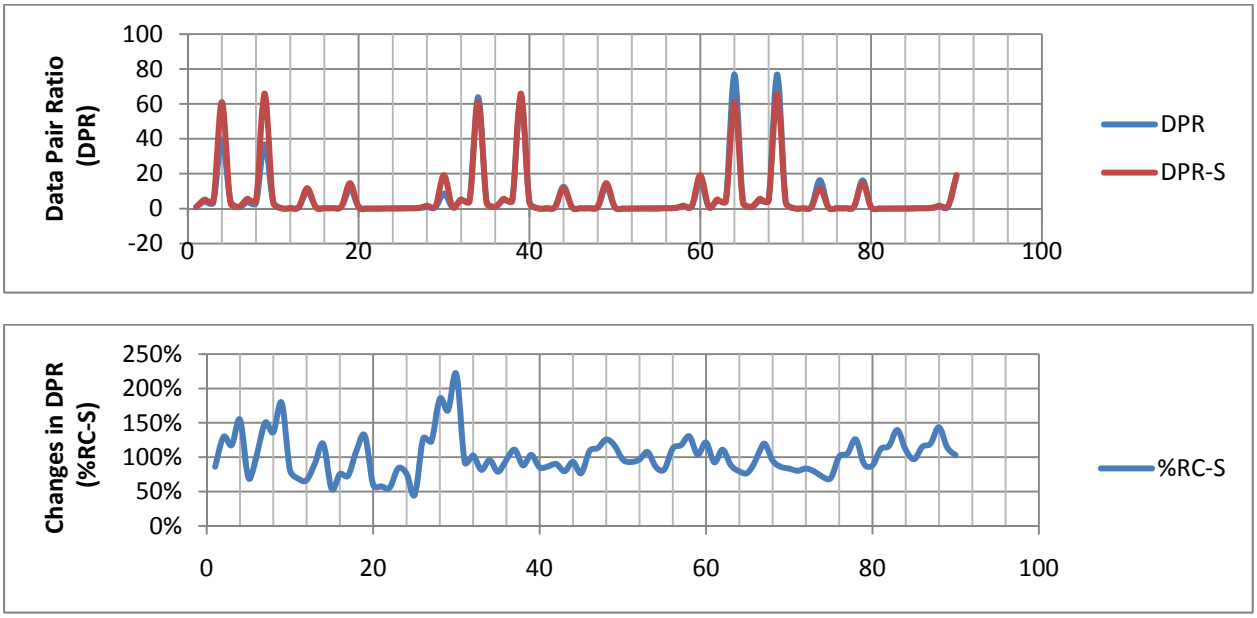




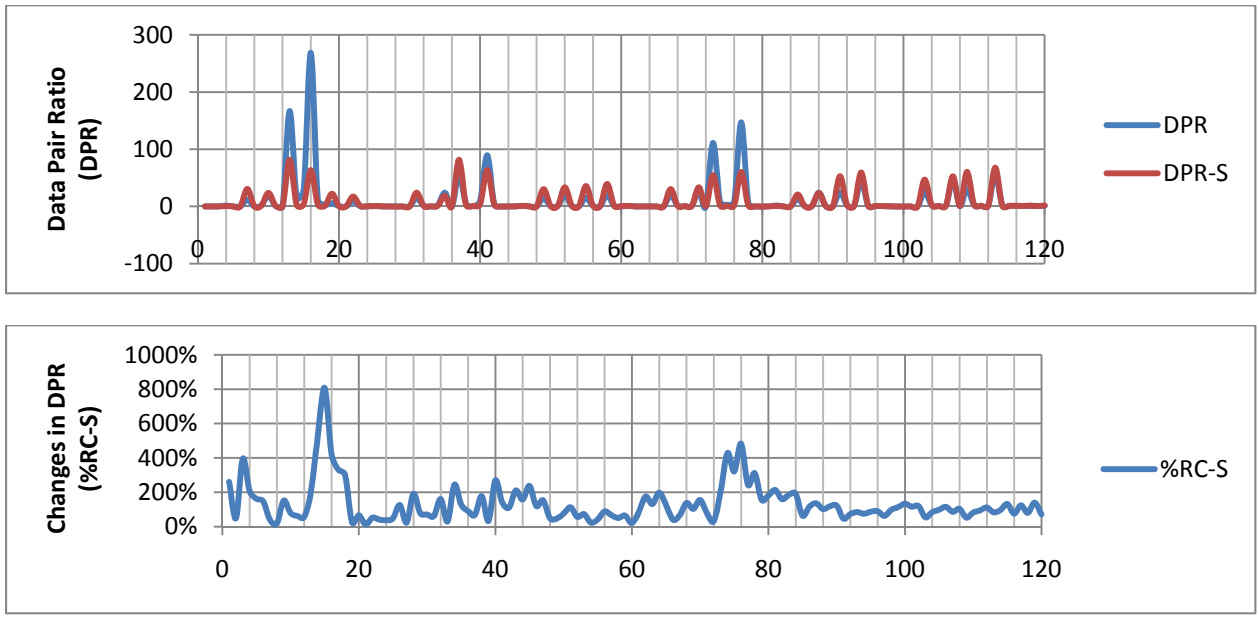
Finally, we can look for patterns specific to schizophrenia by plotting the changes (±15%) as a step chart. This shows us quantitatively what biology did to its connections to transform a normal brain into one with schizophrenia. The run of the step identifies the amount of change for a given data pair (DRE). Notice that prominent runs appear at 0.2, 0.3, 0.4, 0.9, 1, 2, and 3. Such results might be interpreted as a loss of control or the substitution of one set of control mechanisms for another. But what does this tell us about how schizophrenia occurs? It provides a clue. If we can quantify the before (health) and the after (disease), then we can also quantify intermediate way points that may begin to tell us when, where, and how the disease develops.



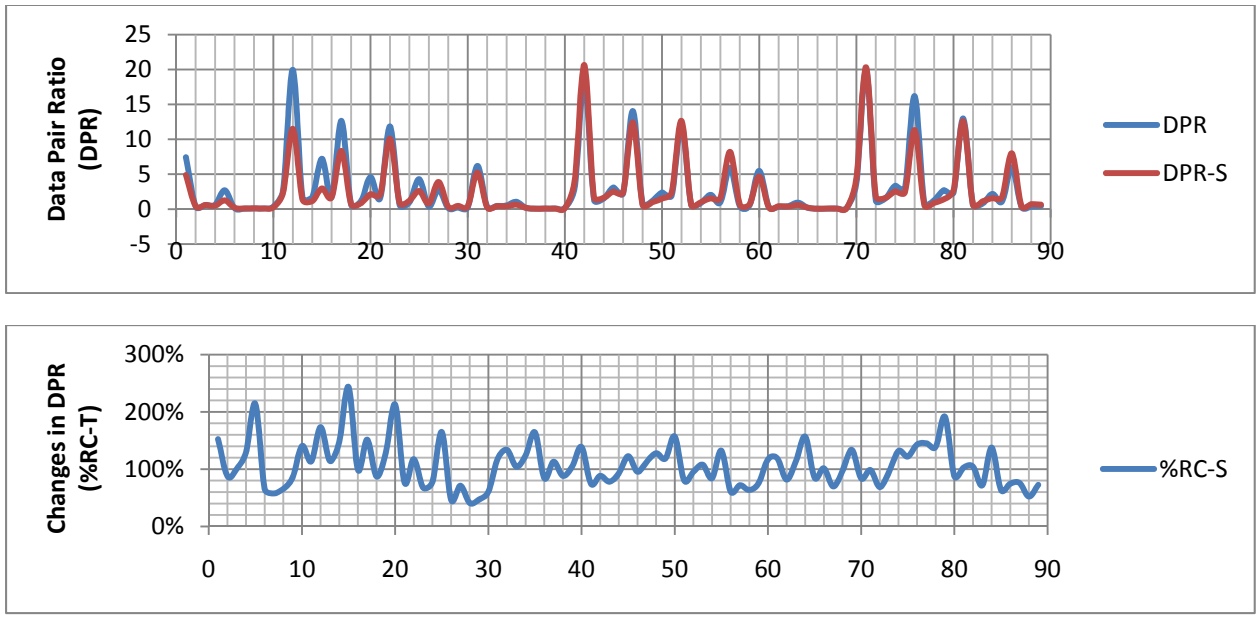
Steady State Change: Diabetes in the rat can cause extensive changes throughout the organism (Cit 4550). The DPR plots compare the kidney, heart, liver, pancreas, and several muscles in health and disease. The % Ratio Changes (%RC-S) flesh out the differences, by comparing controls (DPR) to experimentals (DPR-S).



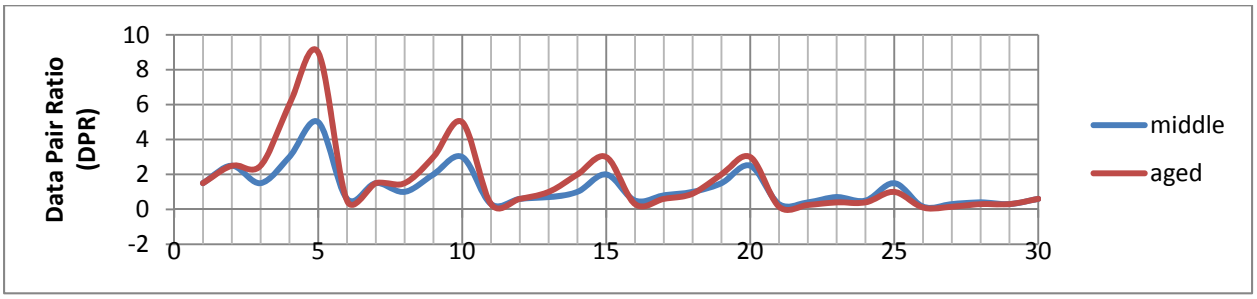
Steady State Change: Aging in rat liver hepatocytes (Cit 1184). Glycogen, Golgi, lipid droplets, lysosomes, mitochondria, and peroxisomes change their relationships over time. The % Ratio Changes (%RC-S) display the differences.

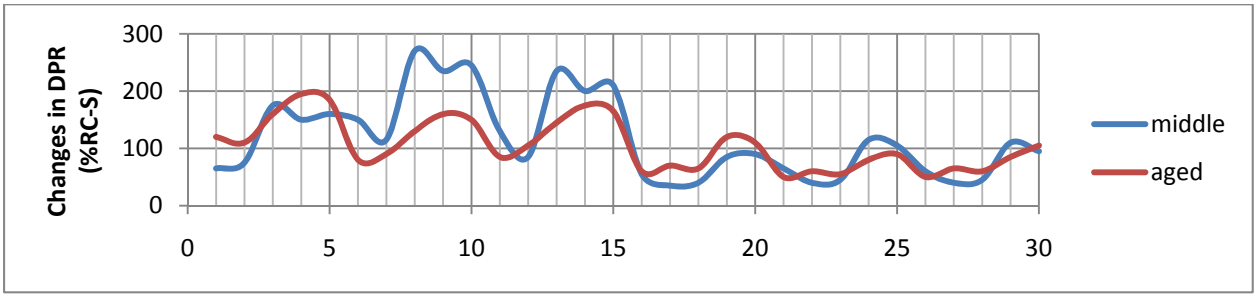


Steady State Change: Effects of nutrition and salt on the major organs of the rat (Cit 5075). Notice how diet and salt intake seem to play important roles in determining phenotypes.

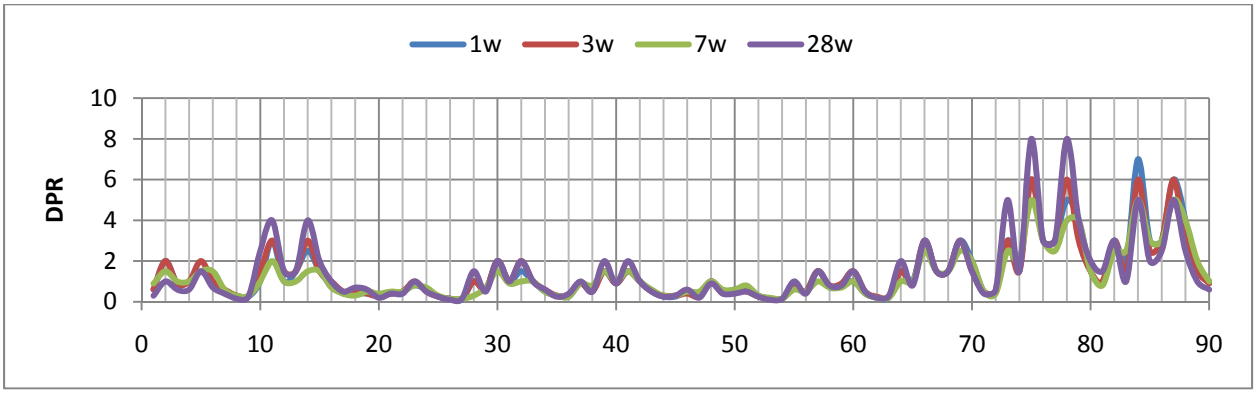


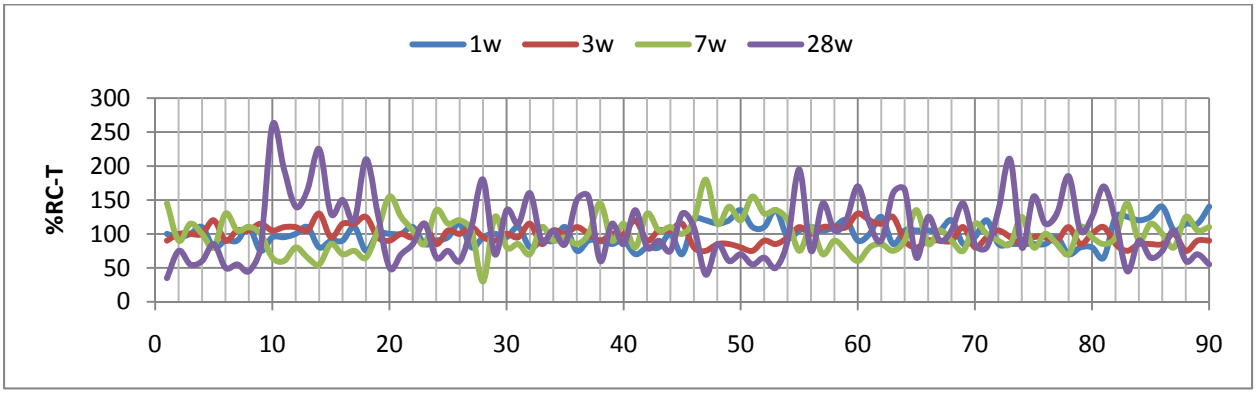
Transitional Change: Aging in different populations of astrocytes of the hippocampus of the rat (Cit 4593). Notice how different populations of astrocytes undergo different amounts of change, which can fluctuate over time.



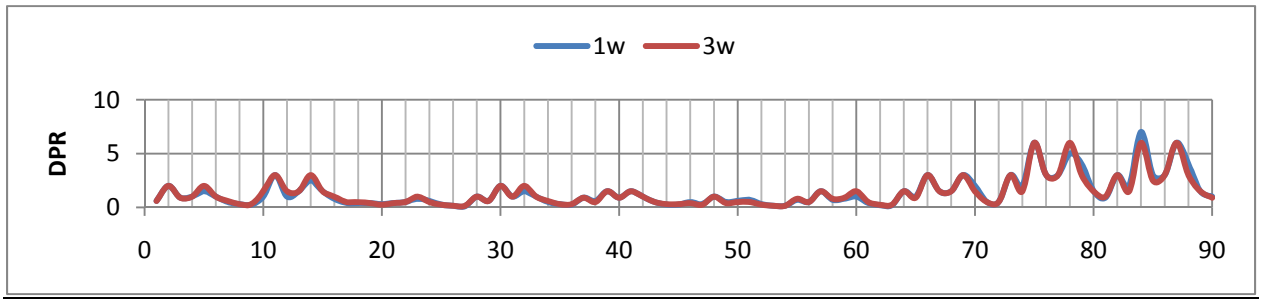


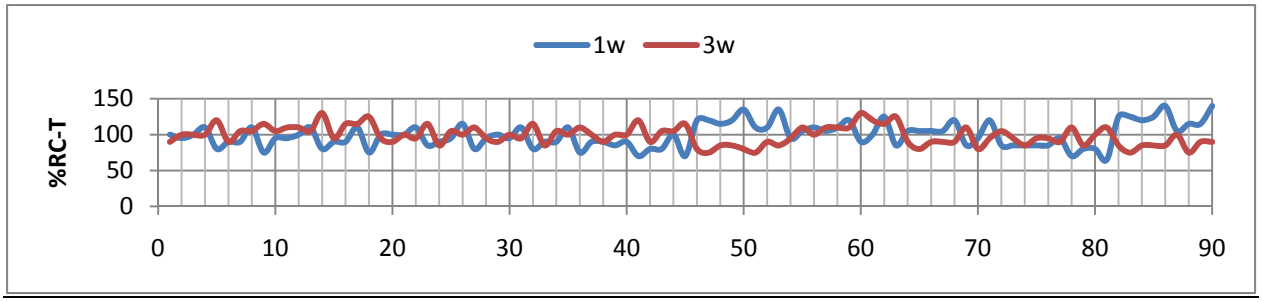
Transitional Change: The effect of inflammation on the cells of the male reproductive system in the rat (Cit 4711) illustrates the remarkable complexity of biological change.



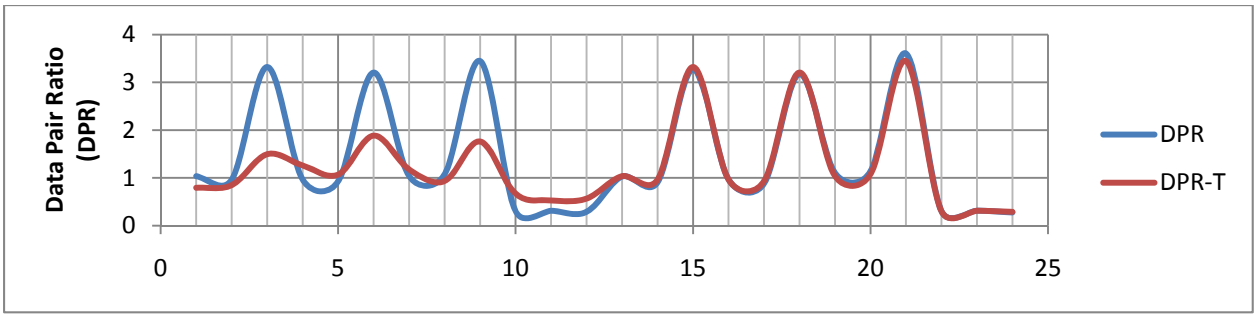


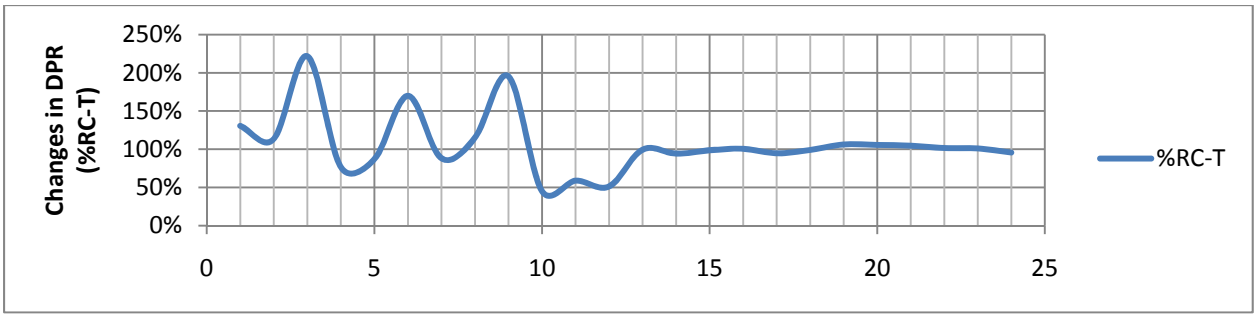
By unfolding the complexities as they occur, the task of seeing what happens locally and globally becomes surprisingly straight forward.



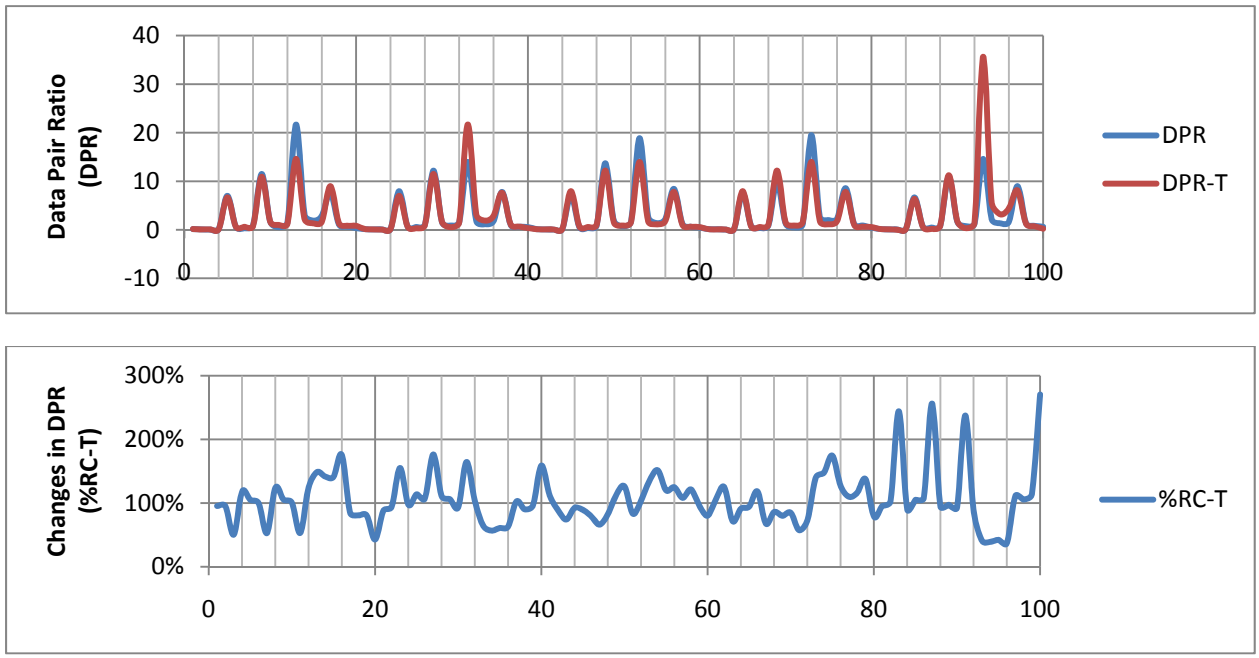


Transitional Change: The effect of aging on neurons in the rat hippocampus (Cit 4331). Notice that for specific populations (marked cytochemically) the number of neurons decreased from young to middle age, but not from middle age to aged.





Transitional Change: The effect of mercury on neurons and glia in the calcarine sulcus of the monkey (Cit 4593). Notice how the patterns change from one time point to the next.



Transitional Change: The effect of adrenalectomy on the hilar, ca1, ca3, and granule cells of the rat hippocampus (Cit 2522).

Notice at 15 and 30 days that the data pairs (DREs) appear identical, as suggested by the red curve sitting directly on top of the blue curve. In contrast, the data pairs of row 4 and 10 suggest slight changes at 120 days.

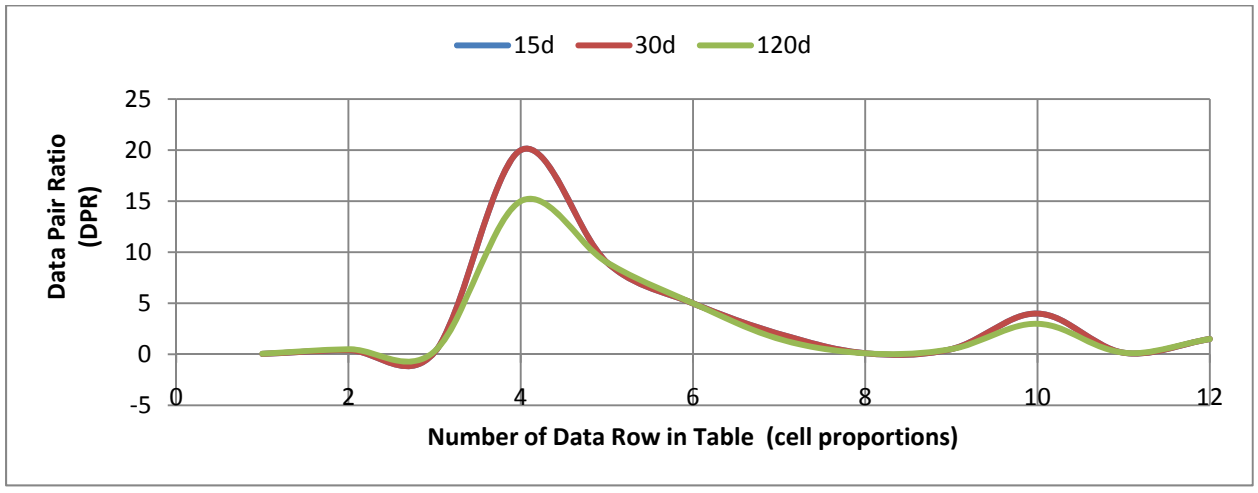


Figure 6 Hippocampus after adrenalectomy. Data pair plots for four cell types - hilar, ca1, ca3, and granule - illustrate the relationship of one cell type to another. See the SB2 folders for the worksheet.

When we look at the % ratio changes in Figure 6, however, a more informative picture emerges. Notice, for example, that the amount of change appears to increase by 30% between days 30 and 120 and that

this change is the same for hilar, ca1, and ca3 cells when compared to granule cells (rows 4, 7, and 10 of the worksheet). Although all these cells types differ in number, they appear to change stoichiometrically as members of a connected set. In effect, they respond as a single unit. Notice too that for these cells the amount of change increased at 15 days, decreased at 30 days, and increased again at 120 days. This oscillating behavior between transitional and steady states often appears during biological changes.

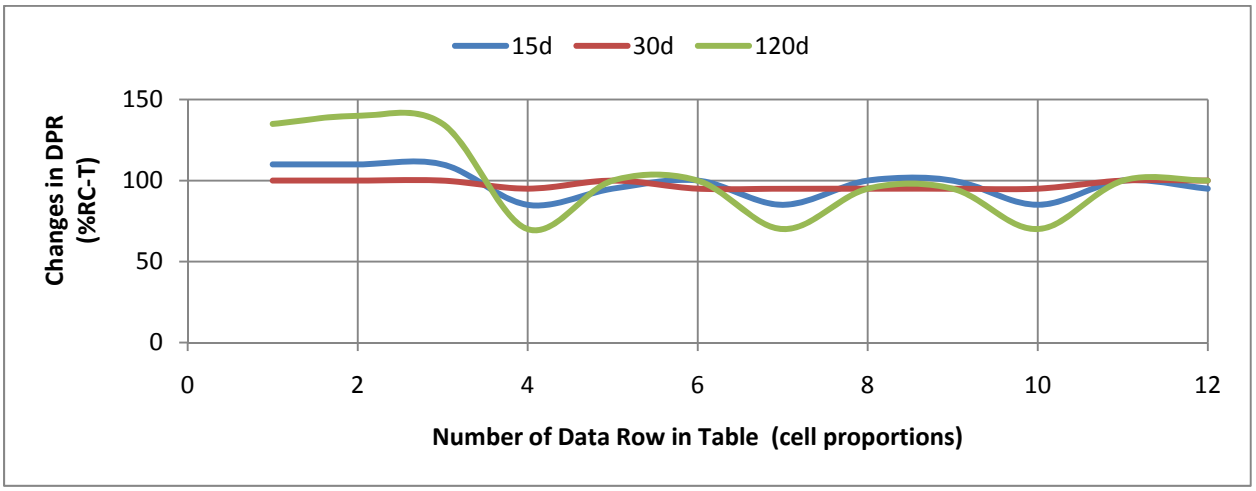


Figure 7 Hippocampus after adrenalectomy. The curves display transitional changes: control->15d; 15d->30d; 30d->120d.

Systems Biology Two

If we set as our first goal a quantitative understanding of how biological phenotypes change in health and disease, then a first step consists of assembling a library of phenotypes for **SB2**. This library now contains folders of worksheets (Excel files) grouped by disease, exposure, and development. As the individual folders fill up with worksheets, we will begin to see what the local and global patterns can tell us how biology engineers its changes.

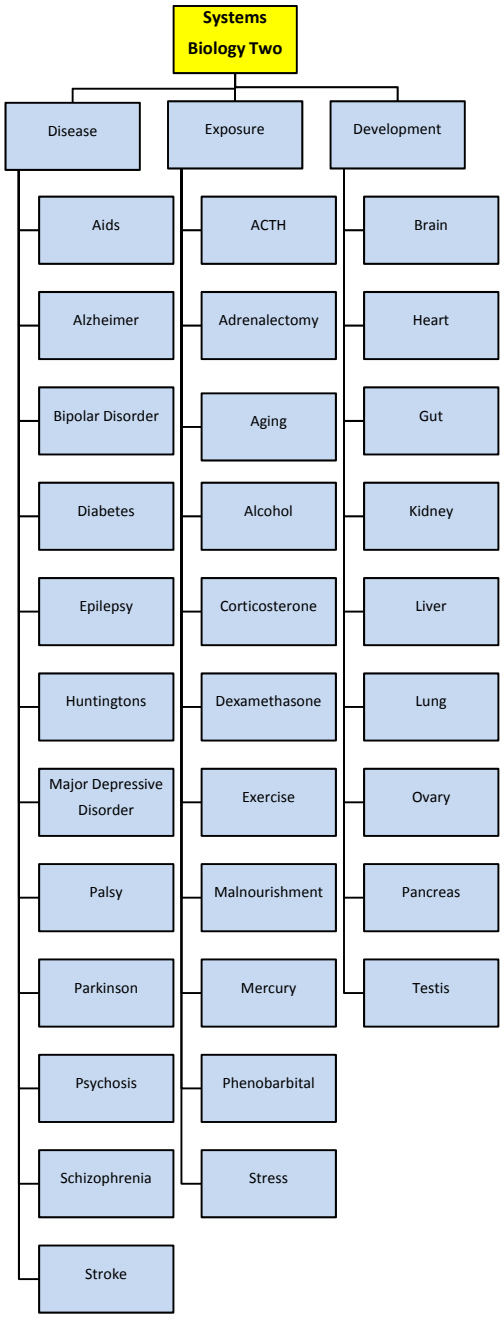


Figure 8 Folders of data and plots for SB2

Discussion

The report and software package suggest a strategy for assembling systems biology from a preexisting information infrastructure. Such an approach seems entirely justified because systems biology, as a byproduct of this infrastructure, immediately reaps the benefits of a thoroughly tested approach to biological complexity.

To be successful, systems biology will have to demonstrate an ability to solve specific problems – particularly those related to health care. This suggests that data sets will tend to be highly focused and may not provide the expected generalizations and predictions. In contrast, an information infrastructure avoids these limitations by including all data that can qualify as being quantitative. As such, it can serve as a general testing platform for biology and spinoff a host of new byproducts as the community continues to innovate and evolve.

Perhaps the most beneficial feature of an information infrastructure comes to us as an ability to optimize outcomes. By unfolding the complexity of biological change, we can see that it includes two inextricable events: changes in the amounts and proportions of parts. If we duplicate this behavior of biology, the data output of our papers increases by at least an order of magnitude. Enter these same data into databases and their productivity continues to grow – now and well into the future. In effect, the optimizing ability of an information infrastructure dramatically increases the value of the biology literature.

Consider the SB2 files. They include a deliberate attempt to begin the long-term process of quantifying disease. If we can generate very specific information from the literature about the progression of a disease and evaluate the efficacy of treatment protocols, then we can also generate new opportunities for pharmaceutical companies wishing to speed the development and testing of their products. In short, we have every reason to expect that a quantitative approach to disease will offer substantial economic incentives.

One of the most pleasant surprises of the information infrastructure comes as an encouragement to ask lively questions. An example will help. Is it logical for biology, built on the quantitative foundations of physics and chemistry, to exist today as a largely descriptive and semiquantitative science? No. The information infrastructure helps us to understand why. Recall that the science of biology relies on the same data types used by physics and chemistry for detecting changes, namely concentrations and amounts. Such data types can work quite well in a reductionist setting, but tend to collapse into chaos when applied to the complex settings of *in vivo* biology. It turns out that this "quantitative discontinuity" between the physical and biological sciences exists largely as a construct, not as a reality. Complexity itself creates a blinding effect. Biological data – **in biology** - exist not as isolated point sources, but rather as dynamic networks of interconnected information. When we collect and quantify data as point sources experimentally, we lose the connectivity and increase the overall complexity by

accumulating the additional burdens of bias, animal variation, and physical disruption. In effect, we increase the complexity of an already complex setting. Our payback takes the form of a descriptive and semiquantitative science. The information infrastructure can effectively reverse this downward spiral with software by minimizing both bias and animal variation and by reconnecting the parts. In so doing, it opens the door to biological complexity and allows us to move on to the next level. Remember, that systems biology will work only if it deals effectively with complexity.

Information Infrastructure

How do we design an information infrastructure for biology using data published in the biology literature? The question creates a paradox, because the answer arrives only after the infrastructure is in place. Moreover, building an infrastructure turns out to be a messy process, one that consists of following clues that all too often lead to dead ends. On the brighter side, one gets to ask most of the controversial questions up front, beginning with: "Are these data any good?" and "Can these data detect changes correctly?" Unless we agree to answer such questions thoughtfully and correctly, building an infrastructure based on published data may become largely a waste of time and resources. Given the difficulty of finding first-rate *in vivo* data in the biology literature, advancing to our next level of understanding may well require systemic changes in the way we collect and analyze our data. In effect, building an information infrastructure triggers a long overdue wake-up call by supplying a strong dose of reality. Fortunately, for everyone, the stereology community has already established an outstanding record of excellence in working with *in vivo* data. Their success can be traced directly to their willingness to solve the sampling problems surrounding biological complexity.

The lessons learned with SB2 should readily transfer to SB1 because the same rules apply to molecules as they do to the larger parts. Semiquantitative data expressed as concentrations can be converted first into data pairs and then into connection phenotypes to look for quantitative changes in the frequency distributions of well-defined collections of parts (Bolender, 2008). Furthermore, molecular data collected as concentrations can be upgraded to absolute amounts with hybrid hierarchy equations or by simply reporting molecular counts per organ or gland. In fact, everything that can be done in SB2 can be done equally well in SB1, but, of course, on a much grander scale.

What - exactly - do we get from an information infrastructure? In short, we get an innovative discovery platform for the basic and clinical sciences, one that optimizes outcomes. It uses technology to work smarter, faster, and more effectively. A carefully designed and thoroughly tested information infrastructure offers all of the following.

- **Transformative:** An information infrastructure transforms a descriptive and qualitative science into a robust and quantitative science, one that accesses the mathematical core of biology, detects biological changes correctly, and operates successfully within the realm of biological complexity.
- **Reality Check:** Building an infrastructure includes taking a hard look at the way the biological sciences operate today, from collecting data to sharing results. Wherever one looks, one finds

compelling evidence of minimization – especially in *in vivo* settings. Sampling methods appear questionable, "quantitative" papers turn out to be semiquantitative, reported changes often fail to detect biological events correctly, numerical data become lost in graphs, and only a small fraction of the informational content of the data contribute to the analysis. Moreover, detecting *in vivo* changes with just concentrations produces vast amounts of misleading results.

- **Recovery:** An information infrastructure allows us to take biology apart and to put it back together, thereby conserving its complexity.
- **Biological Core:** The principal access route to the mathematical core of biology occurs through the proportions of its parts. Stoichiometry defines phenotypes according to rules of connectivity.
- **Research** Model: Reductionism in biology attempts to simplify something that cannot be simplified. In fact, biology defines itself with the complexity emerging from the stoichiometry of its parts.
- **Productivity:** By moving data from paper publications to digital databases, their information content increases by at least an order of magnitude. Flexibility replaces rigidity and activity stagnation. When stored in databases, the value of published data as a research tool increases dramatically.
- **Systems Biology:** To do systems biology, we have to identify sources of unbiased data, to know how to detect biological changes correctly, to manage complexity, and to be quantitative. Anything less will simply not do.
- **Research Data:** There exists two general classes of research data:
    - Biological - Correctly mirrors quantitative events occurring in a biological system
    - Artificial – Represents a measure of something taken from something biological.
- **Discovery Engine:** A digital database begins as a data catalogue that undergoes refinement to become a discovery platform (e.g., Universal Biology Databases, Biology Blueprint, and Connection Phenotypes).
- **Interpretation:** We routinely report only about 10% of the informational content of our research data, ignoring the often more informative 90%.

Discovery and innovation with the Information Infrastructure

What is the principal engine of discovery and innovation basic to an information infrastructure? Recall that both outcomes derive from an ability to create new information from old, where in our case old represents data published in the biology literature. The new information is the data pair, a ratio formed by dividing the amount of one biological part by another. This ratio, which optimizes published data by minimizing bias and animal variability, defines quantitatively the relationship of one part to another as a proportion. It defines the basic unit of connectivity, wherein two parts sort, filter, integrate, and connect according to their numerical values. Indeed, the productivity of the information infrastructure described herein depends largely on its ability to convert amounts (old) into proportions (new). Table 4 summarizes some of these new information products.

Table 4. Primary sources of innovation and new products coming from the data of the Information Infrastructure

New Products	Amounts	Proportions	Reports
Experiments as equations	✓		
Stereology Literature Database	✓		2001
Data standardization (data entry)	✓	✓	2001
Data integration	✓	✓	2001
Design codes	✓	✓	
Minimized bias and animal variation		✓	
Data pairs		✓	
Universal Biology Database		✓	
Decimal repertoire equations		✓	
Engineering (reverse and forward)		✓	
Biological blueprint		✓	
Connection phenotype		✓	2008
Change phenotype		✓	2009
Change – steady state		✓	2009
Change – transitional		✓	2009
Systems Biology Two		✓	2009

Detecting Changes with the Infrastructure

An information infrastructure changes our approach to analyzing biological data largely by adding proportions to our repertoire. Of the options for detecting change, as listed in Table 4, most take advantage of the optimizing properties of proportional data. Moreover, proportions drive discovery and seem ideally suited to the mission critical task of dealing with complexity. As discussed in previous reports, complexity represents a numbers game with regard to data - more being better than less. By assembling data pairs from absolute amounts and densities (concentrations), we immediately benefit from the multiplicative effect of reestablishing biological connections and generating equations. For example, it is now possible to routinely increase the amount of data coming from many publications by at least an order of magnitude. Access to such large data sets becomes a determining factor because only they can supply the robust patterns that ultimately lead to new insights and solutions. Just think for a moment about what we are currently doing. The biology literature, built at a staggering cost over many generations, is being allowed to remain largely fallow, contributing only a tiny fraction of its enormous potential to the scientific community.

Table 3 tells an interesting story. It lists eleven ways for detecting changes, using data within the framework of an information infrastructure. Curiously, only two items on this list appear in most papers being published today – concentrations and absolute amounts. Since concentrations do not count because they routinely fail to detect changes correctly (Bolender, 200) , only one remains - absolute amounts. Outside of the stereology literature, however, such absolute data do not appear that often – the obvious exceptions being body and organ weights. This limitation on detecting change correctly may go a long way in explaining why the productivity of biology continues to fall far short of expectations. Today, biology continues to exist in the Pre-Complexity Age because of our data gathering and interpretation errors and because of our unwillingness to also publish our basic and clinical research data in digital databases.

Everything seemingly reduces to cause and effect. If we cannot detect a change correctly, we cannot interpret our results. If we cannot write equations for our experiments, we cannot know if we are detecting biological changes correctly. If we cannot collect data using unbiased sampling methods, we cannot defend our data or our interpretations. If we cannot become a quantitative science, then how can we expect to meet our obligations as investigators now and in the future?

Health and Disease

Before setting out to fix something, it often helps to know exactly what is broken. A similar logic applies to fixing diseases, particularly chronic ones. Consider what happens. When we cross the line between health and disease, our phenotypes exchange success for failure. Our biological system encounters adversity, fails to deal with it, and begins to decline.

Enter systems biology with its promise to reinvent health and health care. As a new and largely untried discipline, systems biology is in the process of defining itself by establishing a record of accomplishment and productivity. We can contribute to this vetting process by giving it problems to solve and then watch carefully to see how well it solves them. The first problem put to the test consisted of quantifying the successes and failures of phenotypes – in health and disease. The solution required two new data tables for detecting changes (steady state and transitional), quantitative phenotypes, and a library of worked examples (the folders of Systems Biology Two).

What can we learn from this first exercise? Systems Biology Two can be built entirely with data coming from the biology literature. To obtain phenotypes, however, published data have to undergo an activation process. This consists of moving the data through several programs of the information infrastructure: Stereology Literature Database (data entry, standardization) -> Universal Biology Database (data pairs) -> Changes (steady state, transitional) -> connection phenotypes (DPR plots + % Ratio Change). The products, which include these activated data fitted to curves, are beginning to supply new and detailed information about the complex nature of biological change. Noteworthy is the ability of the infrastructure to integrate data coming from one or more papers and to express control and experimental data as standardized and comparable curves (quantitative phenotypes). In effect, this exercise offers a solution to the problem of generating and integrating large-scale data sets from the literature – a capability essential to the task of uncovering the details of biological complexity.

Systems Biology Two

Being successful, you may recall, often depends on creating an environment favorable to such an outcome. Consider the strategy employed herein. By optimizing the biology literature with technology, we now have ready access to the information stored in published data – concentrations, amounts, and proportions. Moreover, an information infrastructure allows us to discover what these data can do for us (Table 4). By doing our homework first, we get to design our systems biology by simply going to the

infrastructure and selecting parts already built and carefully tested. In effect, we increase our chances of a favorable outcome for systems biology by transferring the success of one system to another. Such a strategy puts us in the fortunate position of being able to ask very specific, health-related questions within a well-defined framework.

Systems biology invites a host of questions, many of which we can now begin to answer. Why does biology take such care in maintaining proportions? The answer may be that the proportions play a key role in defining its phenotype. A certain arrangement of parts may produce a beneficial set of emergent properties that remain largely unaffected by background fluctuations in absolute amounts. Perhaps, the phenotype represents an optimal configuration of parts capable of producing the best outcomes. How do diseases resemble or differ from other diseases? Do common patterns of change occur during development? Do cells and tissues apply developmental-like strategies when recovering from drugs and toxins? Does aging diminish the ability of organisms to optimize their phenotypes?

Many more questions spring to mind when viewing the **SB2** files. Notice that given the chance or enough time, perturbed systems tend to return to their original phenotypes. Does this suggest the presence of something biological akin to steady state "software" displaying adaptive behavior? Does this same "software" degrade in disease and aging, as some of the examples in the files suggest? Does the optimizing capability of biology begin to fail over time? Notice that groups of parts often respond together proportionally, even when changes in the magnitude of the absolute amounts differ. Do these groups reveal the presence of local control mechanisms that predict the earlier behavior of molecules and genes?

Concluding Comments

An information infrastructure opens wide the doors to biological complexity. With it, we can now begin that tantalizing journey that continues – uninterruptedly – from organisms to genes. Everything along the way will be quantitative and connected, bound inextricably by mathematics and technology. The single element missing, or at least in very short supply, is an unbiased sampling method for identifying and counting molecules correctly *in vivo*. In time, however, such a method will no doubt appear. Clues may already exist ( ). Given the inescapable role of complexity in biology, software becomes the primary engine of discovery and the literature supplies most of the fuel. Bring both together with technology and return them to the community and we increase our chances of success in moving biology up to the next level of understanding in biology – that of complexity. Building this information infrastructure has been an exercise in optimizing outcomes at every step of the process, from design to testing to products. By using it as a discovery and productivity tool, we can begin to put a new strategy into practice. It now appears likely that biology operates by deciphering two codes simultaneously, one for genes and the other for information. By building our systems biology on a robust quantitative platform, we too may have a turn at deciphering these codes. The genetic code would seem to be at the heart of **Systems Biology One**, whereas the information code exists at the heart of **Systems Biology Two**. If true, then we still need to figure out how two hearts can become one.

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.