In: Enterprise Biology Software, Version 10.0 © 2010 Robert P. Bolender

Enterprise Biology Software: XI. Research (2010)

ROBERT P. BOLENDER

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

SUMMARY

The Enterprise Biology Software Project began – ten years ago – with the assumption that organization is the first step toward understanding complexity. What did we learn? Not only is the assumption correct, but our success in dealing with complexity becomes a direct measure of our success in understanding the quantitative side of biology. If we think of complexity as the game, then we become the players, advancing from one level of complexity to the next. By defining complexity in terms of parts and connections that change, we can readily set our level of play. The entry level includes concentrations, which in an *in vivo* setting detect changes correctly only about 50% of the time. Most published research today operates at this level, wherein reductionism deliberately avoids complexity and all of its rules of play. Stereology takes us to the intermediate level wherein unbiased sampling methods provide estimates for concentrations and amounts that can detect biological changes unambiguously – most of the time. An information infrastructure takes us to an advanced level wherein the data of the biology literature can benefit from the mathematical foundation provided by stereology and from the integrative and pattern-forming powers of relational databases. Now the complexity game includes concentrations, amounts, proportions, and connections, which, when taken together, create an advanced complexity tool – the quantitative phenotype. Such a phenotype is remarkable in that it tells us how many things change simultaneously in complex biological settings. In short, we have learned that the effectiveness of our research becomes largely a function of our level of play. Play at the advanced level, for example, merely requires an understanding of how to assemble new discovery platforms in the form of digital libraries. The current report introduces a new library designed to decipher biological codes that contribute to defining phenotypes quantitatively. It also begins a discussion of the Human Phenome Project. The software package accompanying this report includes a complete set of databases, libraries, software tools, documents, and updates.

INTRODUCTION

Accelerating learning and discovery in the life sciences depends ultimately on our ability to understand biological complexity. This continues to be the principal finding of the Enterprise Biology Software Project. Compelling evidence for such a conclusion comes directly from the process of building and applying an information infrastructure derived from the biology literature. The infrastructure, which creates a gateway to complexity, provides a host of new opportunities for addressing mission critical problems in education, research, health care, and business.

Last year, this infrastructure was assembled into a software package including databases, tools, and documents and distributed with the annual report. The challenge this year is to use it to discover additional properties of the quantitative phenotype.

The central theme of the information infrastructure is one of complexity – how to unfold and refold it locally and globally. Complexity, you will recall, identifies the interactions of many parts, which, in our case, includes parts of many different sizes distributed throughout the biological hierarchy. By finding and quantifying connections among these parts, we can detect quantitative patterns capable of supplying new forms of information. Since we now have an operational information infrastructure, we can use the data contained therein to ask questions of increasing complexity with the expectation of finding credible answers.

Our story this year begins with biology and its ability to store information as codes. Recall that we can define a code as a rule for converting one type of information into a different form or representation. For example, DNA codes for RNA, which, in turn, codes for proteins. This coding process continues downstream as biology forms and organizes everything from individual parts to entire organisms. Indeed, one can readily imagine a grand constellation of codes expressed as rules and extending outward from the genome to all parts of the organism.

The question, of course, is how to test this idea experimentally. Consider this. If the genome codes for the organism, does the organism also code for the genome? In other words, if biology unfolds codes to produce organisms, can we follow these codes back to the DNA – by running the "codes movie" backwards? The answer, of course, becomes yes – if we can crack enough of them to be convincing.

Why would it be useful to the biology community to have access to these organism codes? The short answer comes from the fact that we don't know what we don't know. Currently, we know that only about 1.5% of the genome codes for proteins, with the remaining 98.5% representing largely a terra incognita. By tacitly ignoring the bulk of the DNA, we run the risk of being blindsided by what may turn out to be a vast unidentified store of critical information. Somewhere, somehow, biology must be storing the directions for creating, organizing, and maintaining all the many parts and functions of an organism. In effect, we may have a major problem amounting to a whopping 98.5% hole in our understanding. Can the process of identifying codes operating beyond the genome tell us something about what these unknown portions of the DNA might be doing?

How, exactly, does one go about decoding an organism? It turns out that answering such a question requires the help of an information infrastructure. Recall that we began the process of unfolding the complexity of an organism by transforming published data first into a Universal Biology Database and then into specific digital libraries (see Figure 1). By forming data pairs (X, Y), it quickly became apparent that parts taken two at a time revealed well-defined stoichiometries (X:Y). Summaries in the form of blueprints displayed these quantitative relationships and provided detailed information about the arrangements of parts in biological systems, including the boundary conditions. The seminal finding was that this stoichiometric order was real and existed throughout all the levels of the biological hierarchy.

However, does this coded information extend beyond the data pairs? By forming a new digital library, we can answer this question and advance to the next generation of codes based on triplets. Consider the following. If we sort the data pair table by citation, decimal repertoire equation (DREs), and names (X and Y), we can identify triplets as two different parts sharing the same relationship (proportion) with a third. This requires two separate searches, one on the X names and the other on the Y names. It turns out that triplets occur quite frequently. Moreover, this sorting procedure detects not only triplets, but also relationships up to and including sets with 11 connected parts. This tells us that proportional coding exists and represents a general property of biological systems.

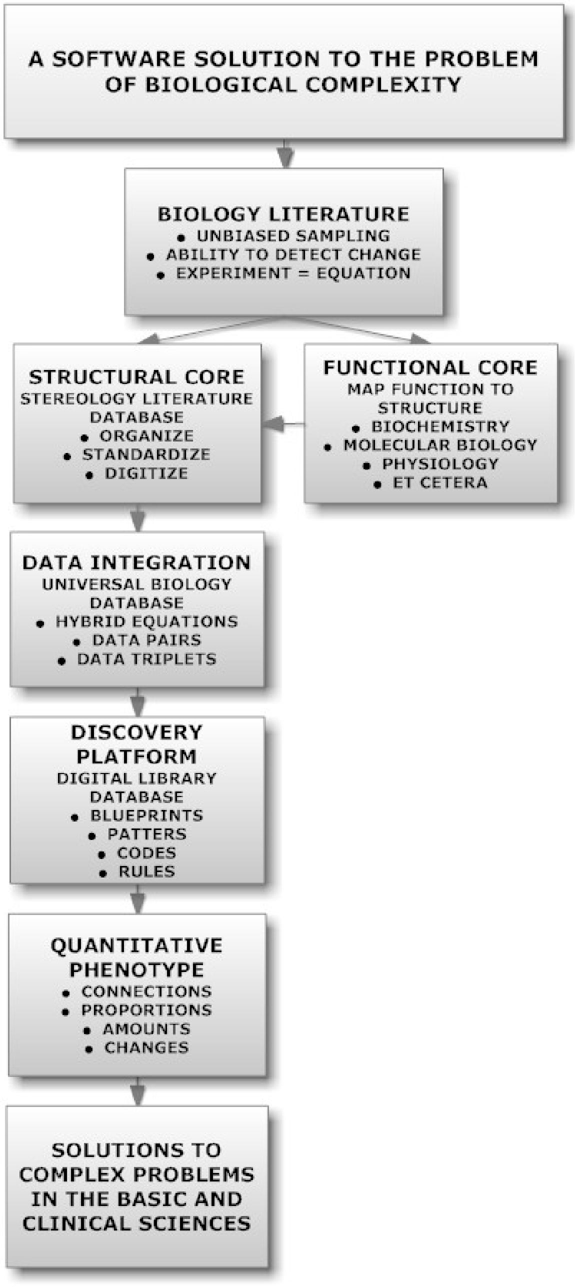


Figure 1. Finding a solution to the problem of biological complexity requires three databases, one to organize all types of biological data, one to integrate such data, and one to serve as an open-ended discovery platform. Complexity – expressed as a quantitative phenotype – becomes the foundation of the solution.

Employing a software approach similar to the one used to form data pairs (X:Y) from the Stereology Literature Database, a new triplet table was prepared and used to generate the next level of code (X:Y:Z). In addition to displaying the usual local and global properties, triplets greatly simplify the task of aggregating this information by allowing the number one to serve as a universal connector. Given a simple set of assembly rules, these triplets can be “snapped” together either visually to see the relationships directly or mathematically to form equations describing a given part, a collection of parts, or a hierarchical network. Note that these dimensionless equations describe proportions capable of reconstituting the original published data (volume, surface, length, number) by simply introducing a single seed value (amount).

By knowing the proportions and amounts of parts over time, we also know the when and where of the production rates. If we can connect these changes in parts - arranged as triplets - to corresponding changes in molecules, then it may become possible to work out explicitly at least some of the local and global rules of genetic expression. Unfortunately, we still do not have a design-based method for counting molecules directly *in vivo*.

In the meantime, however, the triplet library offers a wealth of new information. For example, we can begin to ask fundamental questions about how the proportions of parts contribute to complexity and how complexity is being conserved within and across species. If proportions code for a healthy organism, then what happens to them when a normal environment is challenged by an abnormal one produced by a disease, harmful exposure, or deficit? What responses occur, when, and where?

Curiously, we still know surprisingly little about the complex relationships of proportions to amounts. How, for example, does biology design, use and regulate these relationships? Providing realistic answers to such questions offers a considerable challenge because during a change, parts typically respond at different rates and published data sets often lack identifiable endpoints (e.g., steady states). Nonetheless, the report includes worked examples to illustrate how organism codes allow us to address such complex questions.

What can you - the reader - expect to learn with this new software package? You will become pleasantly surprised by the richness of these organism codes and their remarkable ability to tell us things we didn't know or even suspect. The codes also offer feedback to molecular biology in that they show how large collections of parts routinely change in highly orchestrated ways. In effect, they provide the big picture by showing what happens to the phenotype downstream.

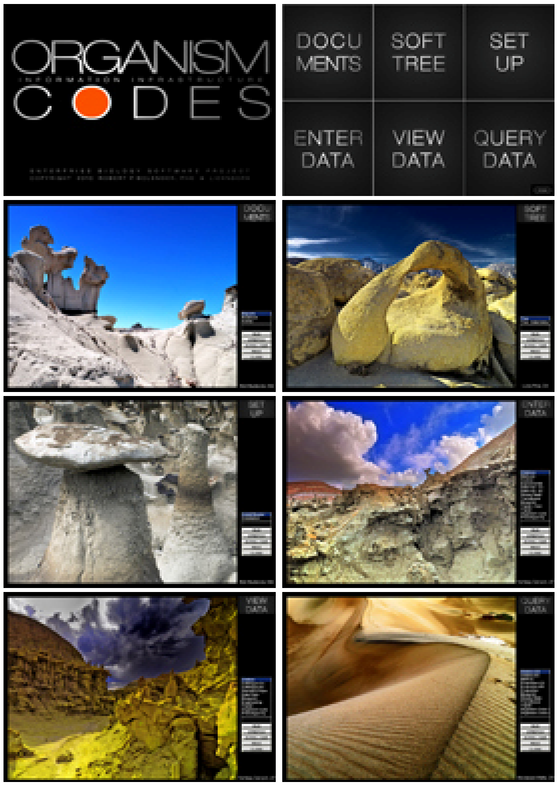


Figure 2. Enterprise Biology Software Package for 2010/2011

METHODS AND RESULTS

The software package for 2010/2011 includes new software tools for unraveling organism codes, which becomes an essential step in understanding the origins of biological complexity.

By creating a new digital library (Organism Codes) within the framework of the Information Infrastructure, the reader quickly discovers the ease with which we can extract new and critical forms of information from our research data.

Enterprise Biology Software Package

The software package includes eight screens offering ready access to programs, databases, and documents (Figure 2). Taken together, they define an information infrastructure based on a mathematics, technology, and data platform.

Information Infrastructure

The task of unfolding biological complexity requires a robust information infrastructure. The flow chart shown in Figure 3 begins with the **Biology Literature** and continues by identifying the steps involved in processing research data into new forms of information. By storing and cataloging these data in relational databases, they become standardized and ready to accept a host of new assignments. Since everything within the information infrastructure remains connected within the framework of a relational model, all information remains readily accessible (Bolender, 2001-2009).

Research in the basic and clinical sciences depends importantly on two basic types of data - amounts and proportions – and on the changes that occur therein. The **Stereology Literature Database** stores amounts and the **Universal**

Biology Database proportions. Since both basic and clinical branches of science continue to rely largely on amounts, proportions remain for the most part a largely untapped resource. Technology, however, quickly levels the data playing field by making these otherwise inaccessible proportions readily available to the research community. This turns out to be a most fortunate outcome because proportions empower the information infrastructure by allowing it to assemble **Digital Libraries**. These libraries become discovery platforms engineered specifically with particular goals in mind. Last year, the report included a library for **Systems Biology Two**, whereas this year it includes one for **Organism Codes**.

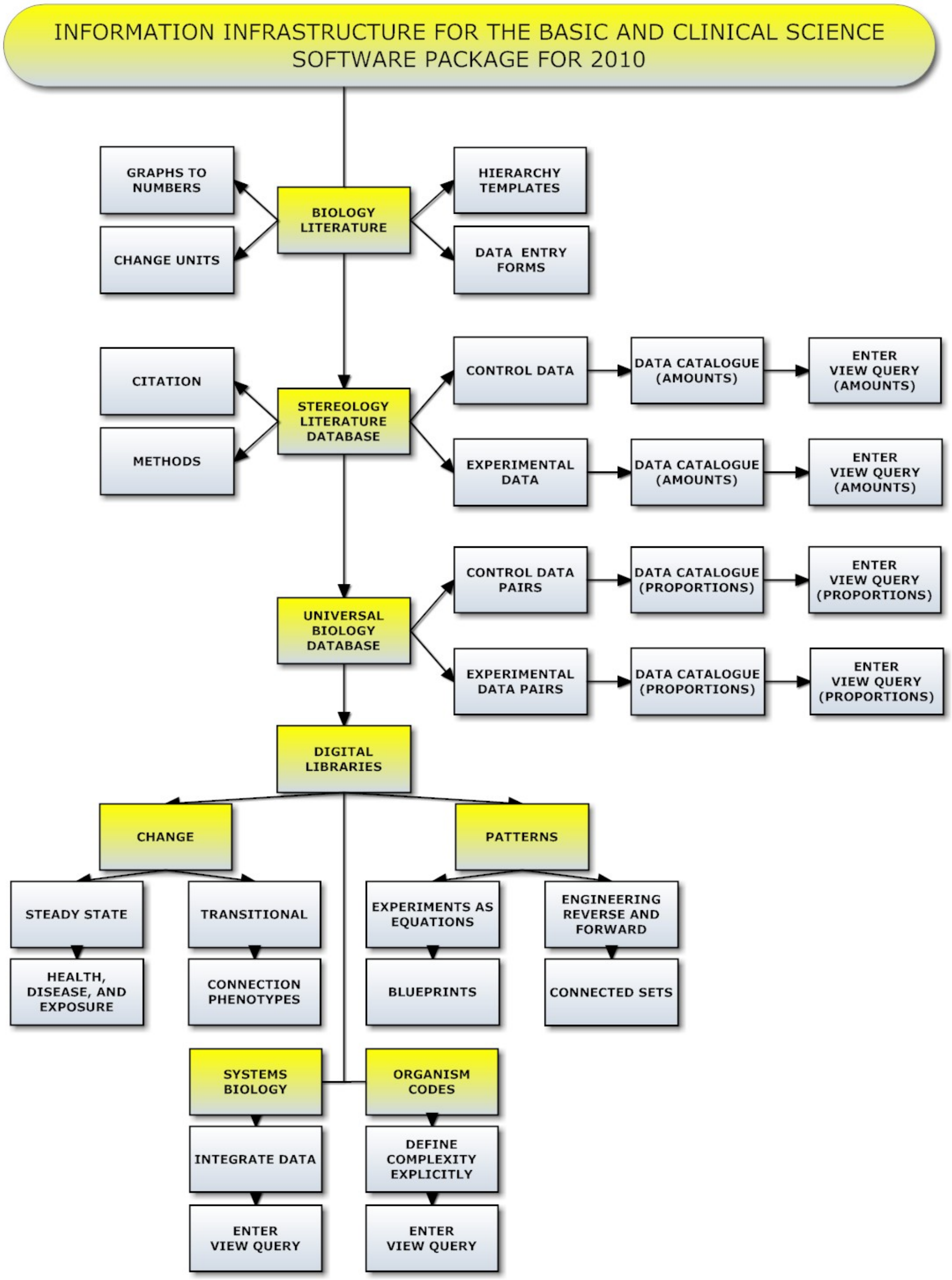


Figure 3. Information infrastructure based on the biology literature.

Organism Codes

Organism codes use published data to detect, diagnose, and predict quantitative patterns in biology. They add another level of information to the quantitative phenotype. We will use them here to consider a fundamental question of information flow. Does information in biology flow in both directions - from small to large (genes to organisms) and large to small (organisms to genes)? If the answer is yes, then we can forward engineer biology from genes and reverse engineer it from organisms. It's simply a matter of choosing a direction of interest. In any case, we already know that biology encodes an enormous amount of information that it decodes as the need arises. Figure 4 expresses this idea visually.

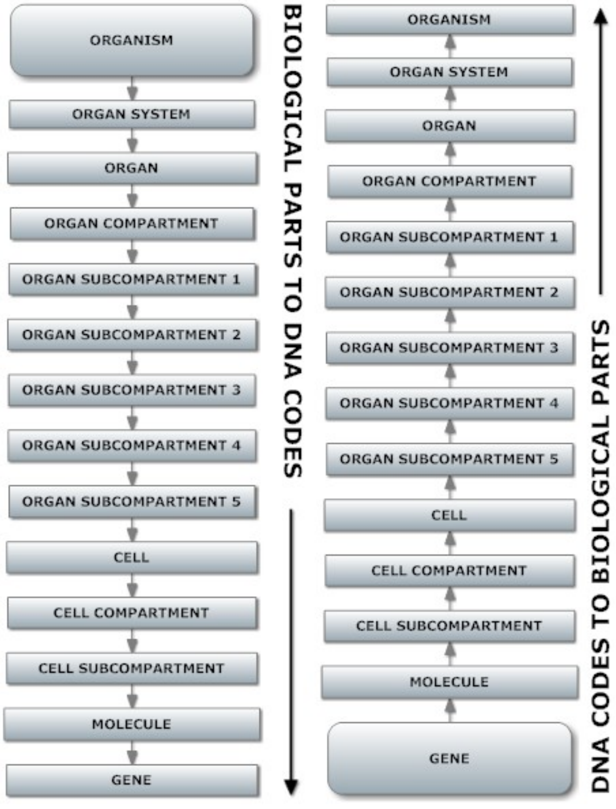


Figure 4. Bidirectional flow of information.

The task before us here will be to figure out how to identify a general solution to the problem of reverse engineering biology from the literature by decoding information currently buried in our published research data.

This story begins with the biological blueprints of earlier reports (Bolender, 2006 to 2009). They showed us that biological parts taken two a time and used to form ratios (data pairs) exhibited highly ordered stoichiometries - distributed throughout the biological hierarchy of size and the animal kingdom. These blueprints provided not only the boundary conditions of nature's design process, but they also told us that biological parts can display different valences, similar to what we see for the elements in the periodic table. In other words, the same pair of parts can occur in different proportions. This represented a critical piece of information because it means that biology employs multiple layers of complexity, but does so in a well-defined and quantitative way. This finding made it possible to reverse and forward engineer the structure of the hippocampus, using published data (Bolender, 2007). Although the data pairs proved adequate to the task of engineering a specific solution, a general solution to the engineering problem remained elusive.

It now appears that a general solution requires little more than upgrading the basic unit of information in the infrastructure from data pairs (i.e., doublets) to data triplets. This process consists of assembling a new digital library called **Organism Codes,** as illustrated in Figure 5. By moving the research data of a given paper through a series of steps, we can extract and visualize an organism code. Since this extraction process includes several detailed steps, a brief example may prove helpful.

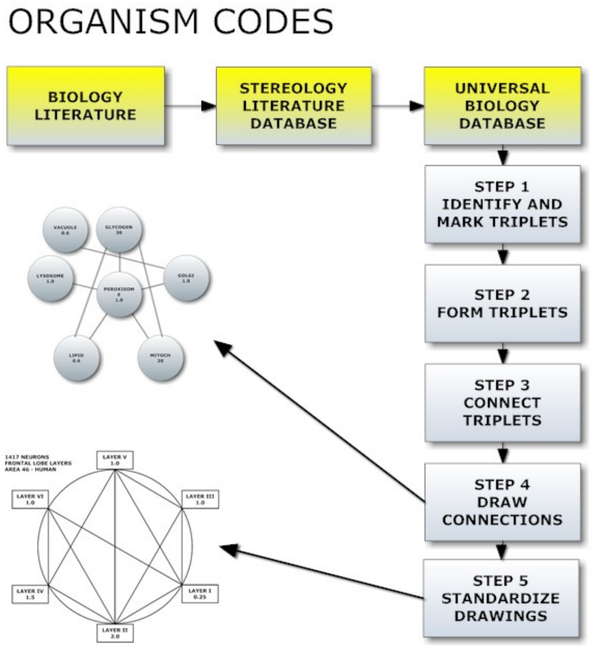


Figure 5. Extracting organism codes from the biology literature.

Step 1 - Identify and mark triplets: Starting with the data pairs of the **Universal Biology Database**, the first program (Valences; Figure 6) uses sorting routines to locate the triplets as two data pairs sharing the same part with the same proportion. When marked with a check, these triplet rows become highlighted in green. To assure the capture of all possible triplets, separate sorts are done on the columns labeled **X Name** (left) and **Y Name** (right), as shown in Figure 6.

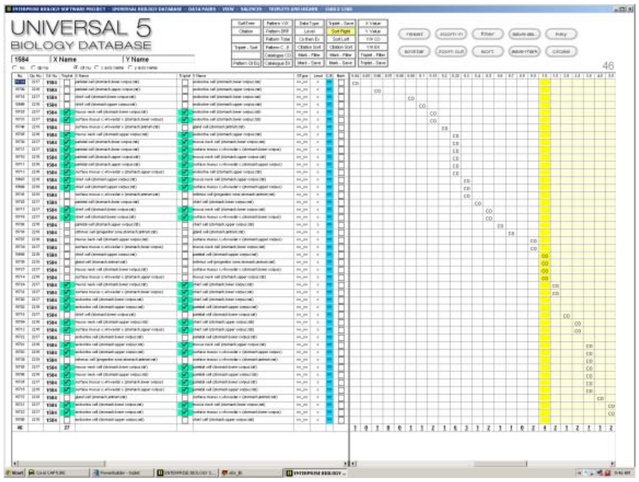


Figure 6. Identifying triplets. Programmed sorting routines (boxed buttons) identify order in the data, one paper at a time.

Step 2 - Form triplets: The next program (Triplets; Figure 7) uses the data pairs identified in Step 1 as triplets (center data window of Figure 7) to form the triplet table. Once again, as done for data pairs, all possible permutations are formed, sorting first on the X Name column (left) and then on the Y name one (right). Programmed buttons automate these sorts.

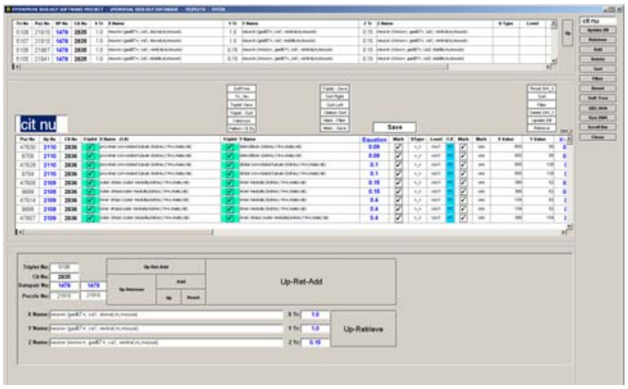


Figure 7. Forming triplets using the data pairs of the Universal Biology Database.

Step 3 – Connect triplets: A third program (Triplets-View; Figure 8), which sorts the triplets by paper, is used to draw the organism codes by hand. This consists of expressing each row of data as three nodes (parts) linked by two lines (connections).

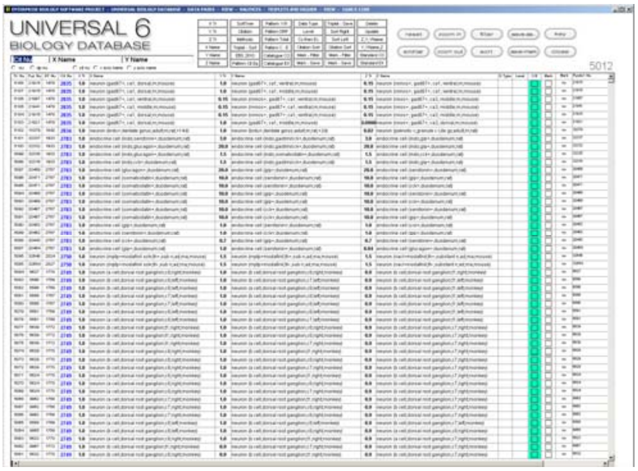


Figure 8. Table of triplets.

Step 4 – Draw connections: The completed sketch becomes the organism code.

Step 5 – Standardize drawings: Finally, for convenience, standardized drawings replace the sketches.

Figure 9 illustrates an organism code for the number of neurons in the human hippocampus. The drawing, which consists of rectangular Boxes (parts expressed as nodes) connected by lines, summarizes the triplets data taken from a single paper (citation 2275).

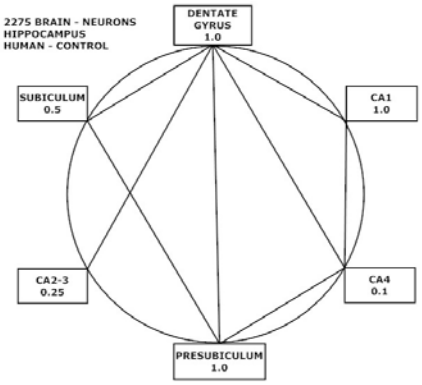


Figure 9. Organism code for the human hippocampus: numbers of neurons expressed as proportions and connections.

Notice that the nodes display the proportions in the number of neurons for the various parts of the hippocampus. Neurons of the dentate gyrus, CA1, and presubiculum occur in similar number (1:1:1), with fewer in CA4 (0.1), CA2-3 (0.25) and subiculum (0.5). Since we know both the proportions and the connections, we also know that a single absolute value will generate values for all the remaining parts. For example, double the number of neurons in the dentate gyrus and all the remaining neurons also double proportionately. In effect, the organism codes offer a general solution to the engineering problem – starting at any point and moving in any direction within the biological hierarchy.

Notice that the organism code equation described below represents a summation:

$$\text{Sum} = \sum_{i=1}^{n} CX_{\text{i}}$$

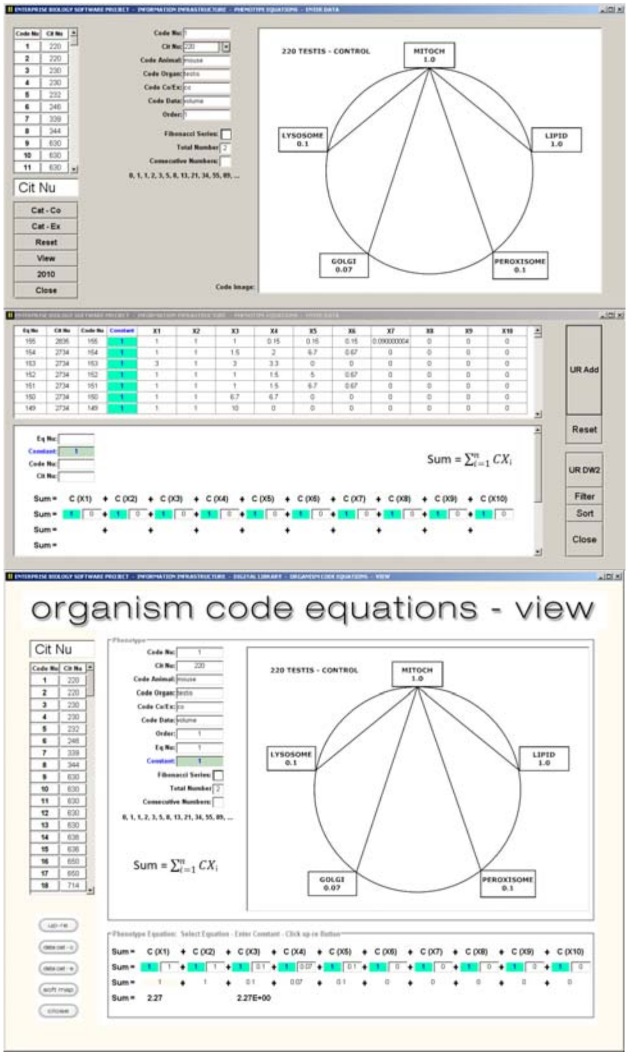
Where C is the seed value and X the variable (proportion). Using an arbitrary number (8,500) as a seed value, the organism code equation for figure 9 is evaluated as follows.

$$\text{Sum} = \sum_{i=1}^{6} 8{,}500X_{\text{i}}$$

Sum = 8,500(1.0) + 8,500(1.0) + 8,500(0.1) + 8,500(1.0) + 8,500(0.25) + 8,500(0.5) = 32,725 neurons

Recall that the data catalogue (Stereology Literature Database) supplies the original published values and the data pair table (Universal Biology Database) includes error estimates for individual data pairs (DREs).

The software package includes screens for entering, viewing, and evaluating organism code equations.


ENTERPRISE BIOLOGY SOFTWARE PROJECT - INFORMATION INFRASTRUCTURE - PHENOTYPE EQUATIONS - ENTER DATA

Code No	Cit No
1	220
2	220
3	230
4	230
5	232
6	248
7	339
8	344
9	630
10	630
11	630

Cit Nu

Code No: 1
Cit No: 220
Code Animal: mouse
Code Organ: testis
Code Co/Ex: co
Code Data: volume
Order: 1
Fibonacci Series: □
Total Number: 2
Consecutive Numbers: □
0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...
Code Image:

Eq No	Cit No	Code No	Constant	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
155	2836	155	1	1	1	1	0.15	0.15	0.15	0.090000004	0	0	0
154	2734	154	1	1	1	1.5	2	6.7	0.67	0	0	0	0
153	2734	153	1	3	1	3	3.3	0	0	0	0	0	0
152	2734	152	1	1	1	1	1.5	5	0.67	0	0	0	0
151	2734	151	1	1	1	1	1.5	6.7	0.67	0	0	0	0
150	2734	150	1	1	1	6.7	6.7	0	0	0	0	0	0
149	2734	149	1	1	1	10	0	0	0	0	0	0	0

Eq No:
Constant: 1
Code No:
Cit No:

$\text{Sum} = \sum_{i=1}^{n} CX_{\text{i}}$

Sum = C (X1) + C (X2) + C (X3) + C (X4) + C (X5) + C (X6) + C (X7) + C (X8) + C (X9) + C (X10)
Sum = 1 0 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0
Sum = + + + + + + + + +
Sum =

organism code equations - view

Cit Nu

Code No	Cit No
1	220
2	220
3	230
4	230
5	232
6	248
7	339
8	344
9	630
10	630
11	630
12	630
13	630
14	636
15	636
16	650
17	650
18	714

Phenotype
Code No: 1
Cit No: 220
Code Animal: mouse
Code Organ: testis
Code Co/Ex: co
Code Data: volume
Order: 1
Eq No: 1
Constant: 1
Fibonacci Series: □
Total Number: 2
Consecutive Numbers: □
0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

$\text{Sum} = \sum_{i=1}^{n} CX_{\text{i}}$

Phenotype Equation: Select Equation - Enter Constant - Click up-re Button
Sum = C (X1) + C (X2) + C (X3) + C (X4) + C (X5) + C (X6) + C (X7) + C (X8) + C (X9) + C (X10)
Sum = 1 1 + 1 1 + 1 0.1 + 1 0.07 + 1 0.1 + 1 0 + 1 0 + 1 0 + 1 0 + 1 0
Sum = 1 + 1 + 0.1 + 0.07 + 0.1 + 0 + 0 + 0 + 0 + 0
Sum = 2.27 2.27E+00

Figure 10. Enter, view, and evaluate organism code equations.

Organism Codes: Local

Organism codes supply both local (one paper) and global (many papers) patterns. We begin with examples of local patterns, which define phenotypes quantitatively in a variety of settings.

Malnourishment: In the absence of a healthy diet, connections between the parts become lost – but for the most part only temporarily.

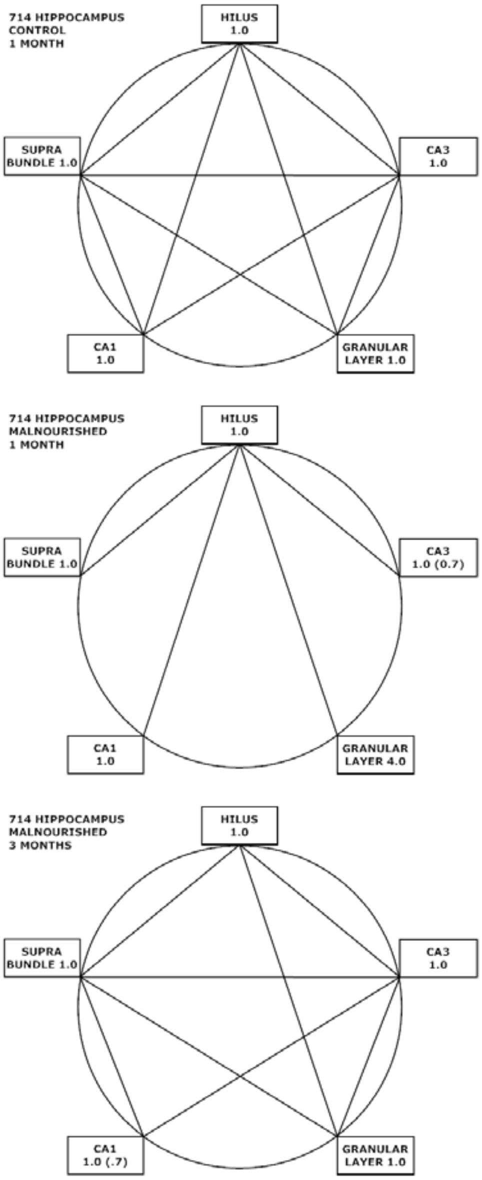


Figure 11. Malnourishment.

Development of the Aorta: Growth and development often occur as cyclic events, alternating between transitional and steady states. We can see this process unfold with the aorta - with an unexpected twist.

The **central organizing structure** – the one to which all or most other parts are connected – begins as the nucleus at day 0, continues at day 2 in the presence of many new connections, and finally reverts to the simple pattern seen at day 0 – except that the organizing structure shifts from the nucleus to the ground substance. As development proceeds, the rules guiding the proportions of parts seem to develop as well. The similarity of the code at days 8 and 12 (not shown) suggests a steady state pause. Note that all six codes are included in the software package and can be viewed – movie-like - by flipping through the screens.

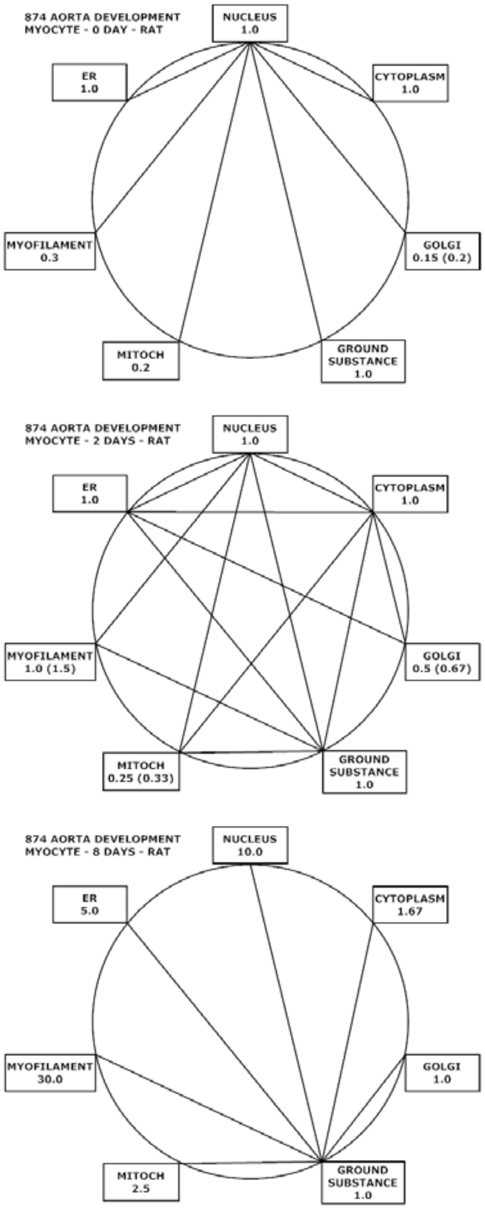


Figure 12. Development of the aorta.

T lymphocyte: Notice how well connected this blood cell appears to be.

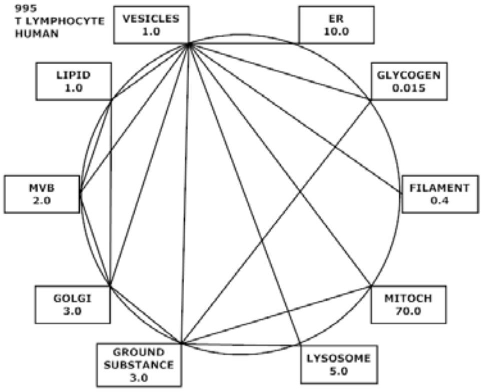


Figure 13. T lymphocyte.

Epidermis of the skin: Cow and rat skin display both similarities and differences.

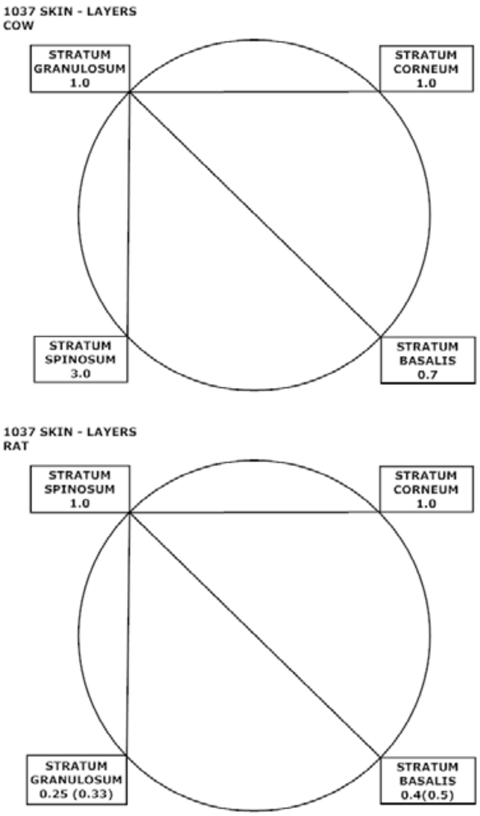


Figure 14. Epidermis of the skin.

Epididymis of the Chicken: Notice the presence of multiple organizing centers, suggesting a redundancy in ordering the relationship of one part to another.

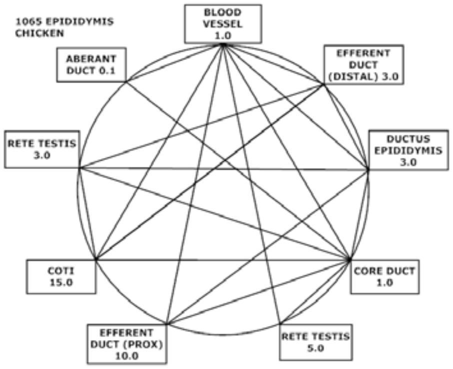


Figure 15. Epididymis of the chicken.

Aging in the Midzonal Hepatocyte: Distinct phenotypes exist for young and old cells, wherein both amounts and connections diminish over time.

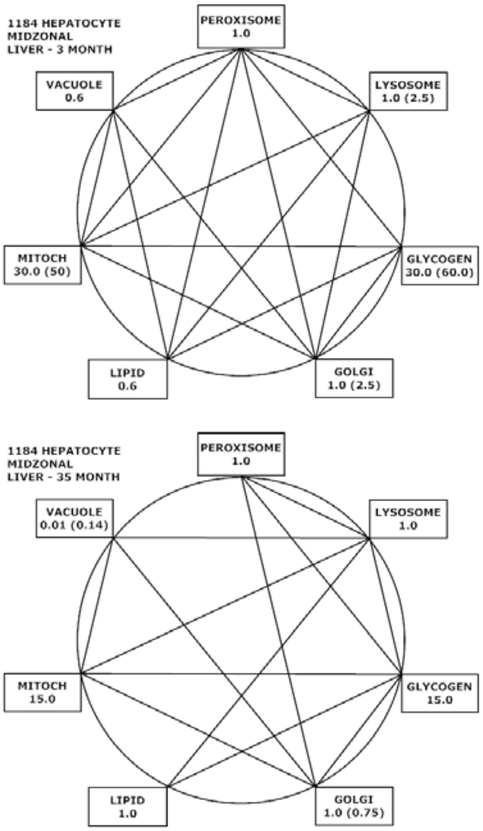


Figure 16. Aging in hepatocytes.

Leydig Cell of the Testis: Cell structure, as defined by the proportions and connections of organelles, appears to be highly species specific – at least for the rat (above) and human (below). What is the implication for animal models?

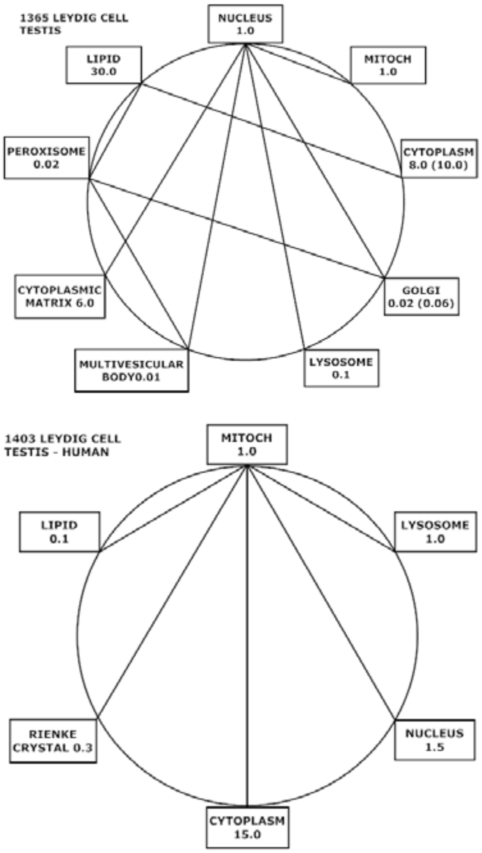
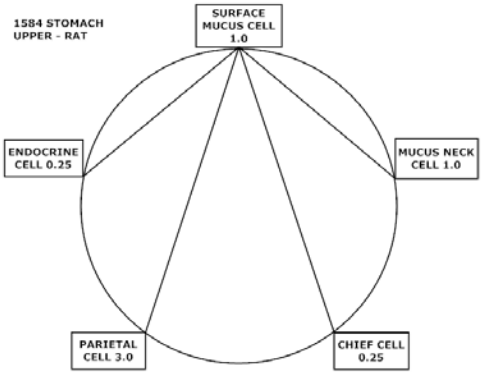


Figure 17. Leydig cells of the testis.

Cells of the Gastric Glands: Different codes exist for the cell populations in the upper and lower portions of the same gland.



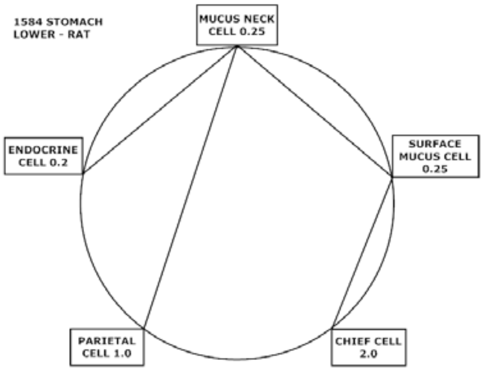


Figure 18. Cells of the gastric glands.

Brain (Adult): Some parts of the brain seem to maximize connectivity, whereas others seem to minimize it. Each part displays a unique and complex set of relationships.

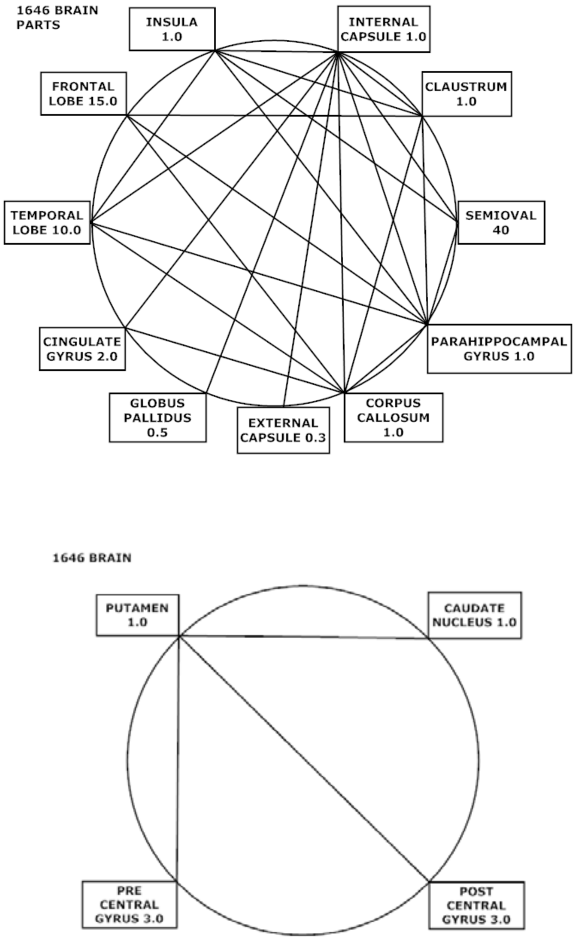


Figure 19. Adult human brain.

Perfusion vs. Immersion: Notice that the parathyroid gland fixed by perfusion displays more nodes and connections than fixed by immersion. Bear in mind, however, that these data come from a single paper, making this a local finding. Before generalizing such a result, however, one would want to inspect the global pattern of fixation (see Figure 33).

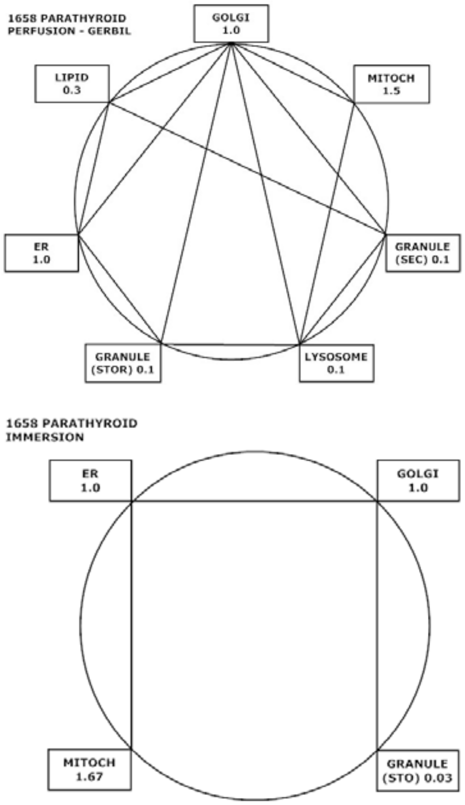
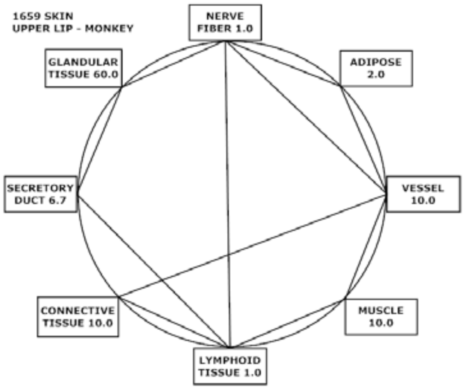


Figure 20. Fixation - perfusion vs. immersion – local data.

Skin of Upper Lip and Soft Palate: The same part (skin) exposed to different surroundings adapts in well-defined ways, as suggested by the following codes.



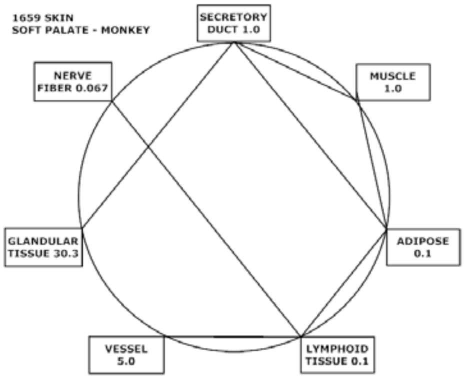
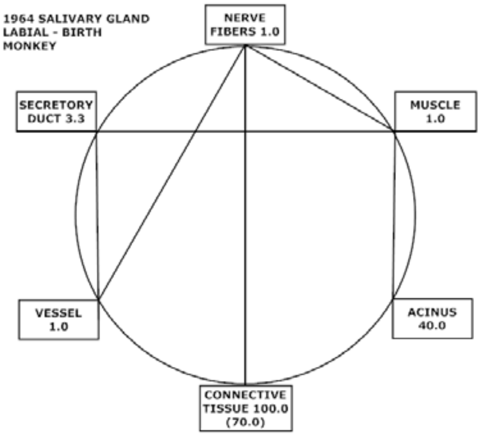
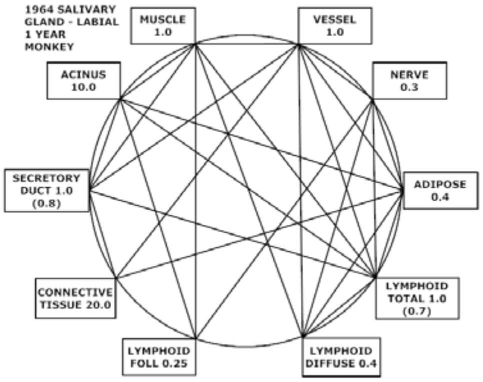


Figure 21. Mucosa of the lip and soft palate.

Aging in the Labial Salivary Gland: Organism codes tend to change continuously over time. The developmental process of assembling parts into a functioning structure apparently requires far more control than maintaining an established one. Notice that the lymphoid tissue assumes its role as an organizing center early (1 year) and persists well into adulthood (9 years). Is this a way to optimize the protective function of this tissue?





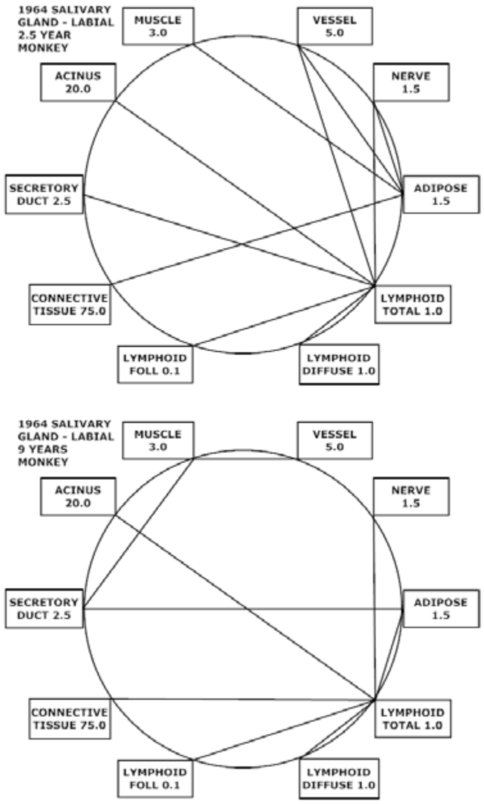
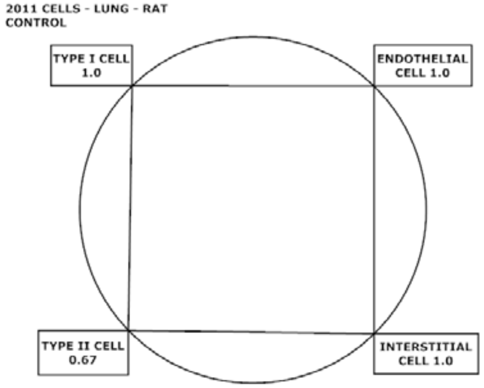


Figure 22. Development of the labial salivary gland.

Radiation and the Rat Lung: The lung responds to radiation by changing the relationships among its cells.



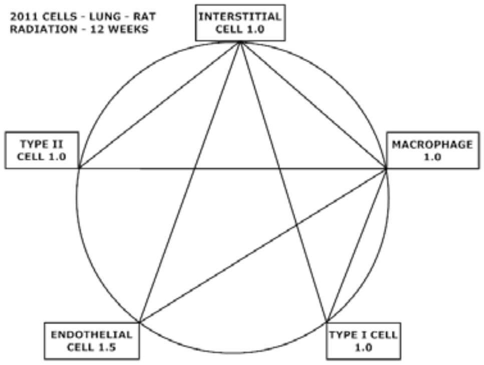


Figure 23. Effects of radiation on the rat lung.

Diabetes and the heart: Disease changes the rules. One code exists for control hearts, another for those with diabetes.

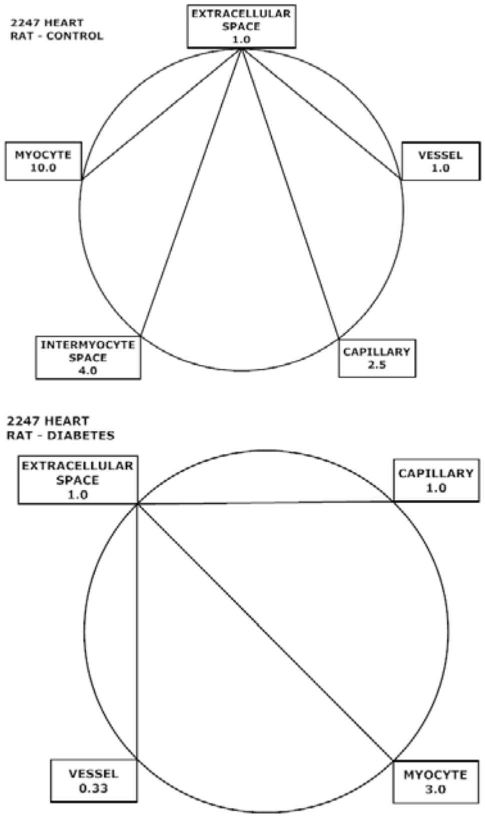


Figure 24. Diabetes and the heart.

Diseases of the Hippocampus: In going from health (normal) to disease (alcoholic, Alzheimer), notice how the dentate gyrus relinquishes its position as a central organizing structure to the presubiculum.

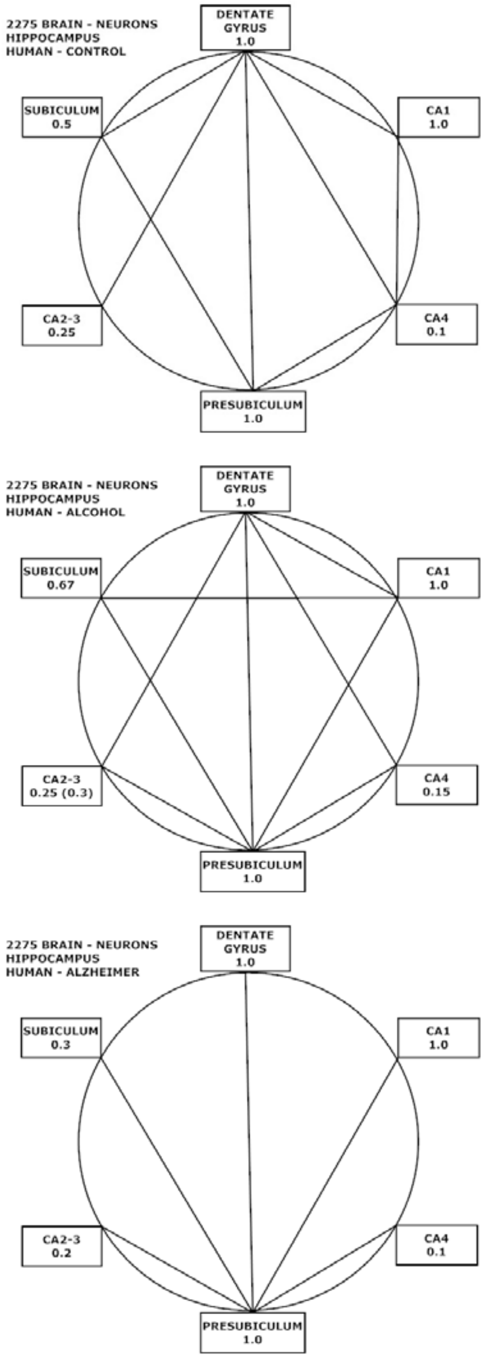


Figure 25. Hippocampus in health and disease.

Kidney Nephrectomy: In response to the surgery, amounts change and the organizing centers shift from the vascular space and inner medulla to the glomerulus. Such changes mark the events of compensatory hypertrophy.

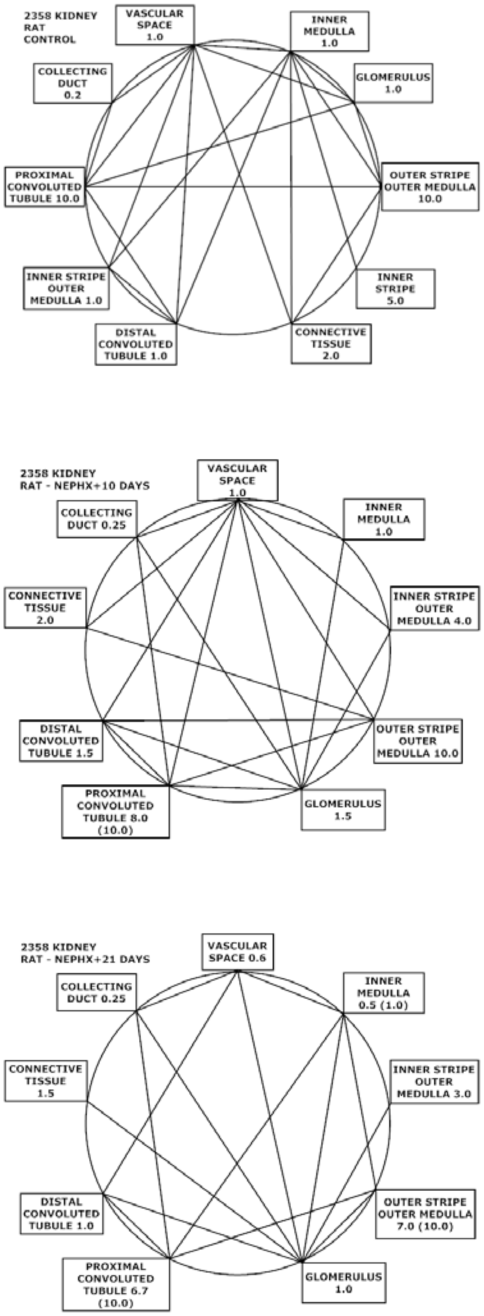

2358 KIDNEY
RAT
CONTROL
VASCULAR SPACE 1.0
INNER MEDULLA 1.0
COLLECTING DUCT 0.2
GLOMERULUS 1.0
PROXIMAL CONVOLUTED TUBULE 10.0
OUTER STRIPE OUTER MEDULLA 10.0
INNER STRIPE OUTER MEDULLA 1.0
INNER STRIPE 5.0
DISTAL CONVOLUTED TUBULE 1.0
CONNECTIVE TISSUE 2.0
2358 KIDNEY
RAT - NEPHX+10 DAYS
VASCULAR SPACE 1.0
COLLECTING DUCT 0.25
INNER MEDULLA 1.0
CONNECTIVE TISSUE 2.0
INNER STRIPE OUTER MEDULLA 4.0
DISTAL CONVOLUTED TUBULE 1.5
OUTER STRIPE OUTER MEDULLA 10.0
PROXIMAL CONVOLUTED TUBULE 8.0 (10.0)
GLOMERULUS 1.5
2358 KIDNEY
RAT - NEPHX+21 DAYS
VASCULAR SPACE 0.6
COLLECTING DUCT 0.25
INNER MEDULLA 0.5 (1.0)
CONNECTIVE TISSUE 1.5
INNER STRIPE OUTER MEDULLA 3.0
DISTAL CONVOLUTED TUBULE 1.0
OUTER STRIPE OUTER MEDULLA 7.0 (10.0)
PROXIMAL CONVOLUTED TUBULE 6.7 (10.0)
GLOMERULUS 1.0

Figure 26. Recovery in the remaining kidney after nephrectomy.

The effect of LH on the Leydig Cell: In response to LH (luteinizing hormone), the nucleus appears in the code, connects to everything, and replaces the ribosome as the central organizing structure.

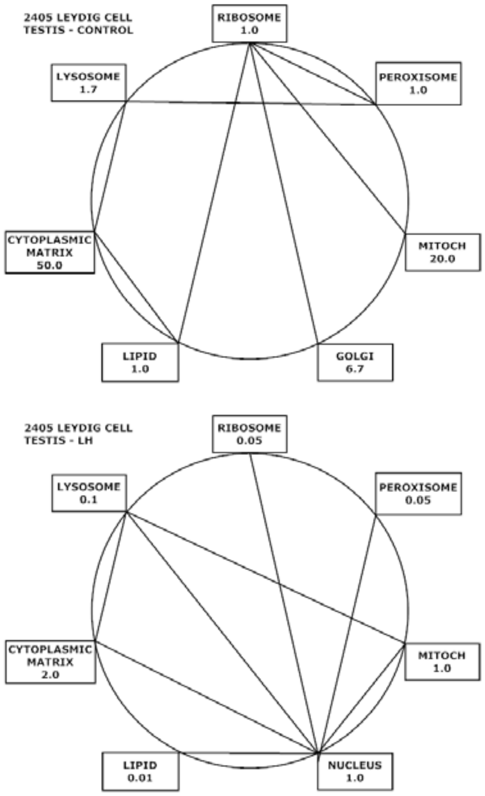


Figure 27. Leydig cells respond to LH.

Adult Testis: In a perpetual state of development, finding many connections in the adult testis comes as no surprise.

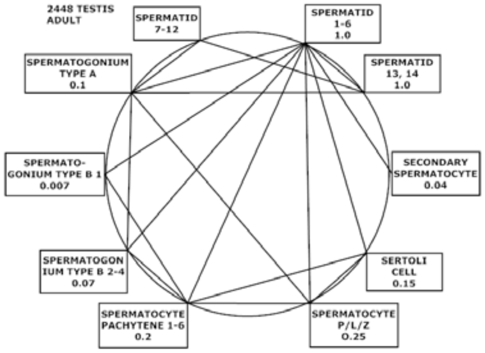

2448 TESTIS
ADULT
SPERMATID 7-12
SPERMATID 1-6 1.0
SPERMATOGONIUM TYPE A 0.1
SPERMATID 13, 14 1.0
SPERMATO-GONIUM TYPE B 1 0.007
SECONDARY SPERMATOCYTE 0.04
SPERMATOGONIUM TYPE B 2-4 0.07
SERTOLI CELL 0.15
SPERMATOCYTE PACHYTENE 1-6 0.2
SPERMATOCYTE P/L/Z 0.25

Figure 28. Adult testis.

Human Brain: Males and females share remarkably similar codes for the volumes of lobes and the numbers of neurons therein.

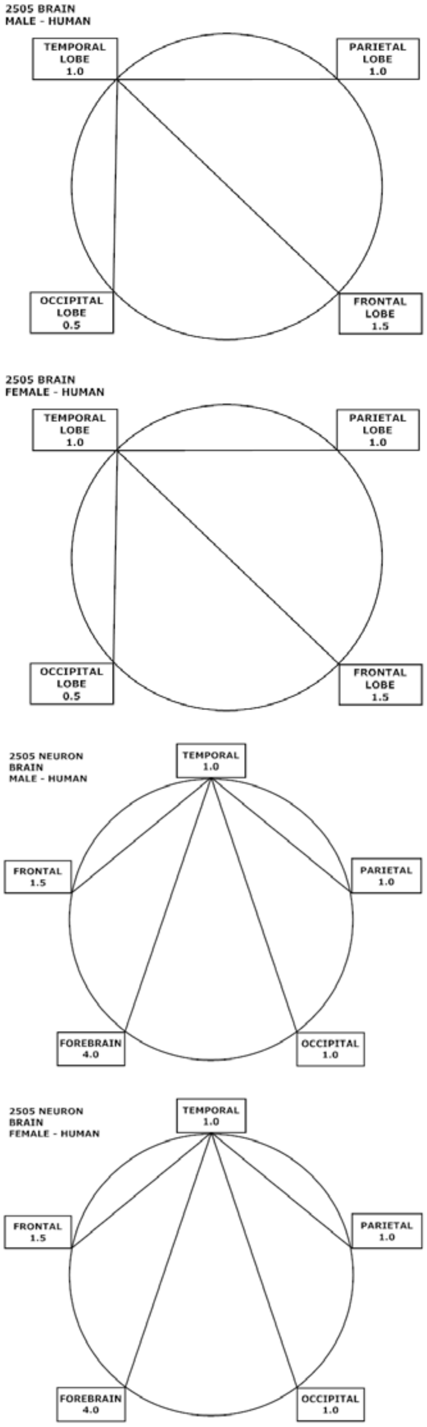


Figure 29. Male vs. female brains.

Testis Development: Notice the distinct patterns of change during development.

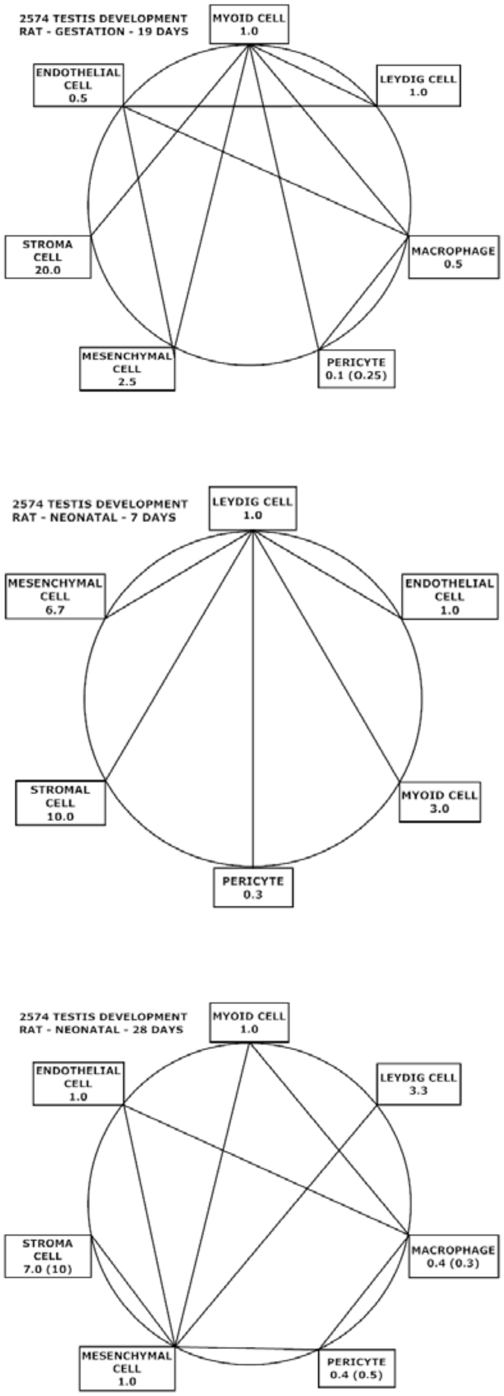


Figure 30. Development of the testis.

Diesel Fumes and Organs: When exposed to an environmental toxin (diesel), organs display a distinct and collective response. They establish connections exclusively to the kidney. Why?

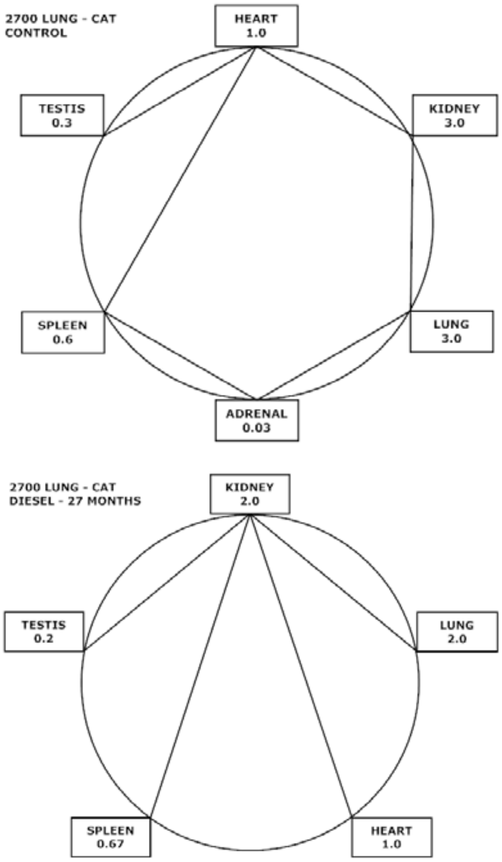
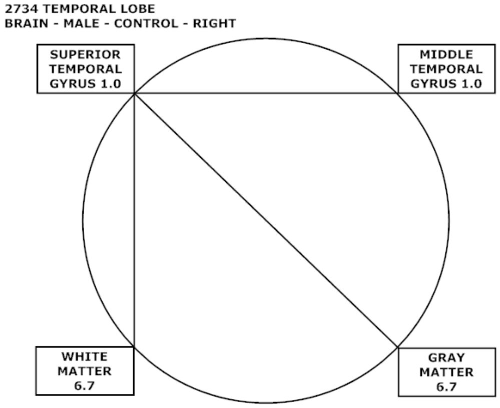


Figure 31. Effect of diesel on major organs.

Schizophrenia: In the temporal lobe, males and females share a similar code in schizophrenia, but not in the normal controls.



2734 TEMPORAL LOBE
BRAIN - MALE - CONTROL - RIGHT
SUPERIOR TEMPORAL GYRUS 1.0
MIDDLE TEMPORAL GYRUS 1.0
GRAY MATTER 6.7
WHITE MATTER 6.7

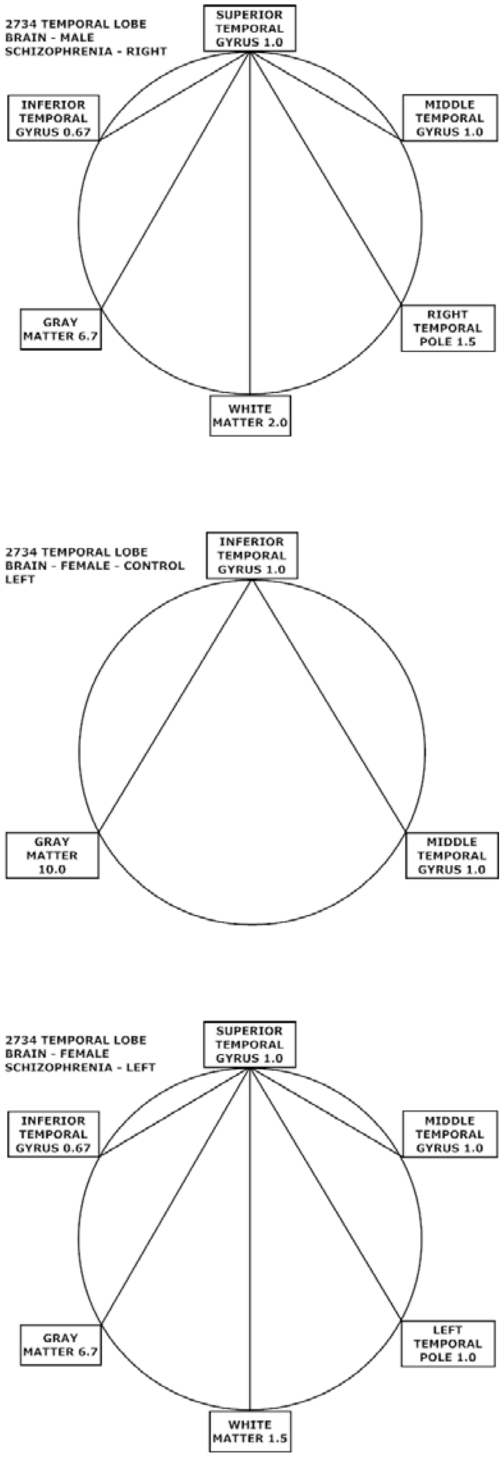


Figure 32. Schizophrenia - male vs. female.

Organism Codes: Global

One of the most remarkable finding of the Biology Blueprint was that the same data pair in the same proportion appeared so many times within and across species (Bolender, 2006-09). This suggests that the relationship of one part to another was being highly conserved.

With a data set only one fiftieth as large, looking for a similar pattern with the triplets becomes somewhat speculative. Nonetheless, preliminary observations may offer us some hints of what to expect. Do triplets with the same parts in the proportions occur in similar and different animals? The preliminary answer is yes. Run the View Triplets program.

Perfusion vs. immersion: Data coming from a given paper can only characterize a given subject under a given set of experimental conditions. The organism codes of Figure 20 offer such an example. These local codes might lead us to believe that fixation by perfusion is the better method because it detected more nodes and more connections. Is this the case? Using the information infrastructure, we can repeat the experiment using all the papers in the organism codes library with fixation data. Now we find both outcomes. Sometimes perfusion appears to win, other times immersion. The point is that understanding often requires global data (Figure 33).

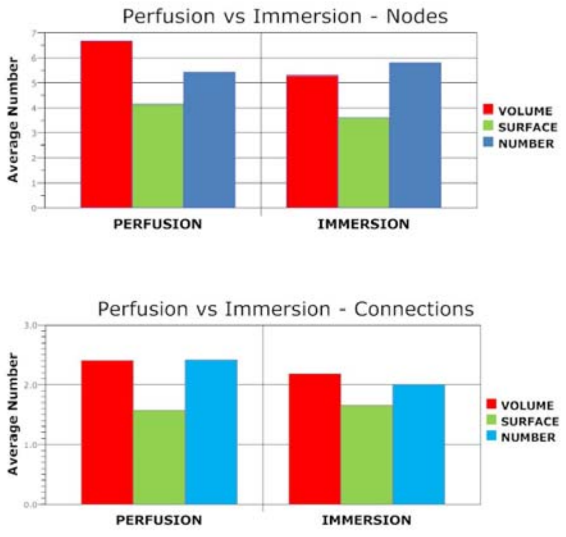


Figure 33. Fixation - perfusion vs. immersion – global data.

Fibonacci Numbers: One of the most mischievous things about biology is that it often embeds codes within codes, thereby increasing the extent of its complexity. The Fibonacci Series (expressed as numbers and ratios), which represents one of the best-known and well-documented quantitative patterns in biology, offers an interesting example. Fibonacci numbers have been identified, for example, in the skeletal system, in the arrangement of biological parts, in defining facial beauty, in animal breeding, et cetera. Do these numbers also apply to the organism codes? Recall that absolute amounts code to data pairs, data pairs to data triplets, and triplets to organism codes. Do the nodes of the organism codes fit a Fibonacci pattern? If we consider the range of the data (nodes numbered 1 to 10), then we can assume that 50% of the nodes will be Fibonacci numbers (1,2,3,5,8,…,n) by chance alone. Here our experiment consists of determining – paper by paper - the proportion of Fibonacci numbers occurring in the nodes – either as total numbers or as numbers following the sequence of a Fibonacci series (1,2,3,5,8,…,n). Figure 34 displays the results.

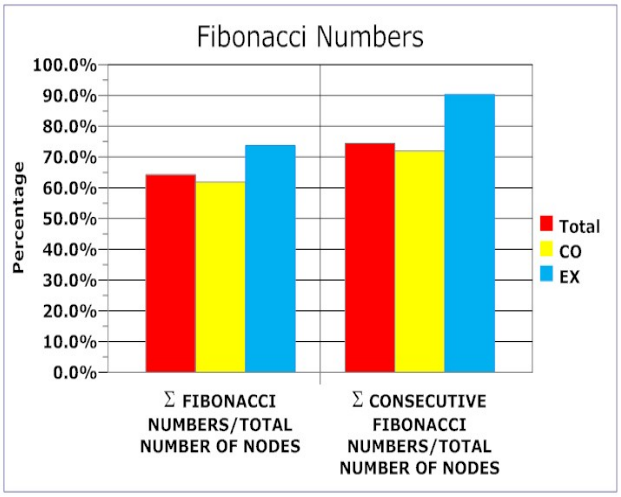


Figure 34. Nodes with Fibonacci numbers expressed as a percentage of the total number of nodes.

Notice that in all cases the percentage of Fibonacci numbers exceeded 50%. The incidence was highest in experimental settings (74%, 90%) when the organism is in the process of responding to a change in its environment. It will be of interest to see if this pattern holds with a larger data set.

Dominant Central Organizers: The largest number of connections often goes to a single node (part), thereby identifying it as the central organizing structure. Figure 35 identifies parts displaying this dominant behavior, wherein the nucleus and mitochondrion hold the top positions. How, for example, might molecular biology explain this observation? Does that fact that they both contain DNA provide a clue?

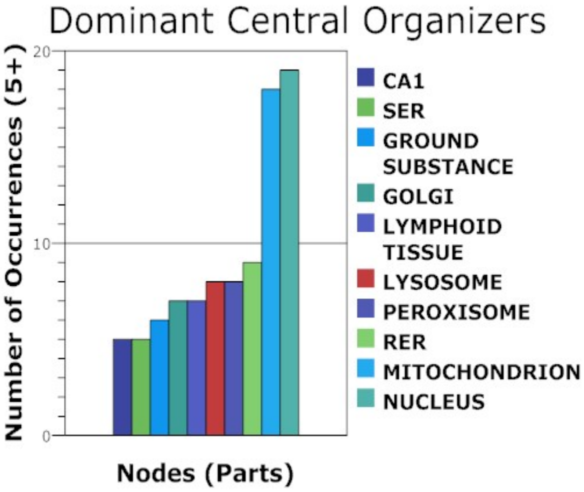


Figure 35. Central organizers include nodes (parts) with the largest number of connections.

New Information

Each new digital library comes with the expectation of new discoveries. The **Organism Codes Library** is no exception. At the top of its list sits the observation that biology orders it parts with a well-defined and mathematical order, an explanation for which becomes a new challenge. Does this order exist in our DNA as a hard coded template, is it the product of an optimizing algorithm, or is it the result of a self-organizing principle?

Table 1 includes a preliminary listing of what the new library seems to be telling us.

Table 1. Properties of Organism Codes

Organism Codes	
1.	Organism codes can be viewed as nodes (parts) connected by lines (proportions).
2.	Some nodes serve as central organizing structures in that they provide connections to all or most of the remaining parts.
3.	An organism code defines an equation connecting a well-defined set of parts.
4.	Codes display a stoichiometry consistent with the presence of an underlying biological principle.
5.	Codes may or may not change when moving from or to transitional or steady states.
6.	Changes in absolute amounts may or may not produce changes in codes.
7.	Codes may or may not return to the original state after a change.
8.	Disease, treatments, and exposures usually alter normal codes.
9.	Experimental methods can affect the detection of codes.
10.	Codes may be conserved within and across species.
11.	Codes can include as many as eleven connected parts (nodes).
12.	Triplets, which make up the codes, can be extracted in large numbers from published data.
13.	Organism code equations predict the amounts of all parts therein from a single seed value.
14.	Codes connect with other codes - within and across all levels of the biological hierarchy.
15.	Codes define phenotypes quantitatively in health and disease.
16.	Codes can define biological complexity as a set of mathematically related parts.

DICUSSION

Since the report this year marks the tenth anniversary of the Enterprise Biology Software Project, a brief look back – and then ahead – seems appropriate. In 2001, the promise of the mission statement included “exploring the future of biology with mathematics and technology” with the goal of “accelerating learning and discovery.”

Retrospective

The inspiration for the project has its roots in the **grand challenge** presented at The Matrix of Biological Knowledge Workshop (Morowitz and Smith, 1987), which defined a new technology model for research in biology as:

“The complete database of published experiments, structured by the laws, empirical generalizations, and physical foundations of biology and connected by all the interspecific transfers of information.”

Reading between the lines, we can see that the challenge became one of reconciling the complexity of the literature with that of the biology it attempts to explain. The practical solution included organizing the literature by storing published data in a relational database and then using technology to unravel complexity as a way of identifying laws, principles, and connections.

Designing a single database for the entire biology literature becomes a frustratingly difficult task because of the counterintuitive nature of the design process. One does not know – a priori – that the design comes not from the designer, but from biology - by way of the literature. To work successfully, the design of the database must mirror both the biology and the literature – simultaneously – a challenging task indeed. Since the purpose of such a database is to explain biology with generalizations, laws, and connections, it requires data collected with unbiased sampling methods and procedures that can detect biological changes unambiguously. Therefore, the design and production of the Stereology Literature Database satisfied successfully the database requirement of the workshop recommendation – minus the word complete.

Extracting laws, empirical generalizations, and physical foundations from the biology literature require – as a minimum - a robust mathematical approach to biology, one that generalizes research data quantitatively across a highly diverse set of publications. The current foundation of biology, however, rests almost entirely on data in the form of concentrations and amounts, both of which turn out to be ill suited to the task of finding patterns because of the distorting influences of animal variation, methodological biases, and uncontrolled complexity. Forming data pairs minimized two of these distortions (by allowing them to cancel out), thereby leading to the establishment of a Universal Biology Database. In turn, this new database became the quantitative platform from which to assemble digital libraries capable of extracting laws, empirical generalizations, and connections from the biology literature. By controlling complexity, the application of advanced technologies allows us to address many elements of the grand challenge successfully.

Taken together the databases, software tools, and methods of the Enterprise Biology Software Project combine to form an Information Infrastructure for the basic and clinical sciences. The central role of this infrastructure is to serve as a discovery platform, one that operates largely by assembling digital libraries.

Figures 1 and 3 illustrate the process of assembling an information infrastructure from the biology literature wherein the whole can become greater than the sum of its parts. As such, the infrastructure becomes an emergent property of the biology literature.

Prospective

Going forward, we now have the wherewithal to map complexity across the organism going from the largest part to the smallest, or from the smallest to the largest. This capability identifies a new grand challenge. How and where does biology store the designs for all its parts – large and small? Does everything come originally from the genes or does the DNA also contain hard-coded templates? Can, for example, DNA also code for organs and organ systems or can it only code for proteins and the timed regulation thereof? How do parts interact with one another and with the genome to become distinct phenotypes?

Organism Codes

Organism codes tell us how specific parts – having many different sizes and occurring throughout the biological hierarchy - relate to one another in a quantitative way. Taken together, these codes capture the organization framework of an organism as a connected set of proportions. This connectivity allows us to go anywhere we please in an organism – quantitatively - simply by moving through the proportions of the connected codes. Consider the implication. Organism codes for example, can be linked into massive networks that can provide a foundation suitable for developing large-scale simulations. Figure 36 illustrates this arrangement as a connected stack of codes arranged hierarchically.

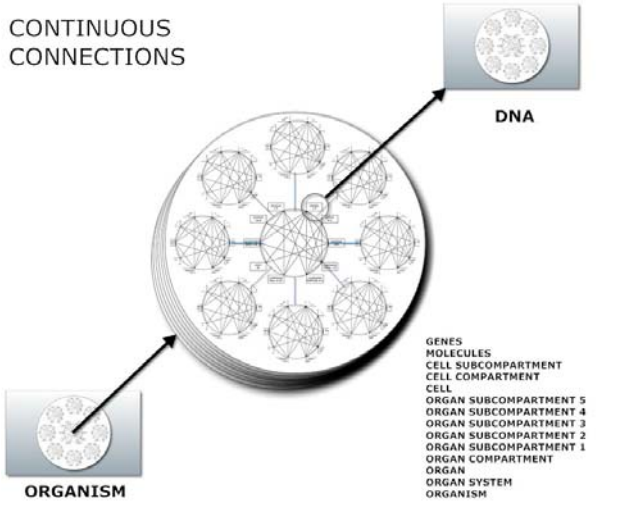


Figure 36. Quantitative continuity.

Observe that by moving up the stack - from the organism to the DNA – we can identify causes, whereas moving down the stack yields effects.

Clearly, the organism codes represent a work in progress in that only about 20% of the codes in the triplet database have been analyzed. However, the codes clearly exist and add another layer of detail to the quantitative phenotype.

Research Data

One of the most unexpected discoveries of the Enterprise Biology Software Project is the extent to which our published data become more remarkable and informative when they become part of an information infrastructure. Such an infrastructure offers a welcome addition to the biology community, one that actively encourages creativity, discovery, and productivity – outcomes essential to our success going forward.

Operating within the framework of an information infrastructure becomes a wonderfully reassuring experience in that we know what we are doing and can understand why. Indeed, our level of comfort increases dramatically. Instead of being lost in the overwhelming complexity created by the ambiguities of semiquantitative methods and data, we can operate comfortably and confidently within the domain of a robust quantitative science. In effect, the information infrastructure quickly transforms biology into a quantitative discipline.

The prescription remains simple and straightforward. Collect data with unbiased sampling methods, detect changes correctly by designing experiments as equations, record data in databases, and report results as amounts and proportions. The information infrastructure does the rest.

Complexity Embedded in Complexity

By looking at biology as a system of codes embedded in codes, we can begin the process of unraveling complexity – step-by-step. The organism codes tell us that biology creates and maintains its parts inventory by throwing a large net of connections over both local and global sets of nodes. This apparent redundancy may help to explain the remarkable resiliency of living systems and their ability to adapt. Although some nodes enjoy dominance as organizers (Figure 35), this dominance often shifts temporarily during change. Such plasticity in the phenotype suggests that biology is either looking up or assembling alternative codes *de novo* as the need arises.

Fibonacci numbers were included in the report because they often appear in the design of living systems and can play a role in self-assembly (Douady and Couder, 1996). In the analysis of the organism codes, Fibonacci numbers appear in frequencies greater than would be expected by chance, becoming most prominent when an organism is in the process of remodeling its phenotype (Figure 34). At such a time, self-assembly may be playing a pivotal role in catalyzing – or in directing - the formation of a new organism code. In such a setting, one can imagine the presence of multiple rules operating simultaneously and in tandem.

The organism codes identify a general property of living systems as an ability to revert to an original or control phenotype. This directs our attention to yet another enigmatic question. How does a specific cell – or for that matter any biological part - recover its phenotypic identity after a temporary change, considering that all somatic cells of a given subject share the same genome? Do phenotypes develop or contain “memories” in the form of template codes that define its preferred configuration? Since phenotypes can effectively “reboot” to a previous design with such remarkable precision, it becomes difficult to dismiss the existence of a built in control mechanism. Indeed, the recent work of Günesdogan et al. (2010) suggests histones as a viable candidate for such a template. They describe – in fruit flies - the presence of an "epigenetic histone code." This suggests that cells can become imprinted in response to a physical location or to a given set of circumstances. It also suggests that turning complexity on and off may be yet another example of a cell’s ability to optimize outcomes.

Track Record

What happens when we reboot the biology enterprise using a new operating system, one that reinvents the biology literature with mathematics and technology? We get a quantitative biology within the framework of an information infrastructure that when combined they become a new engine of discovery, innovation, and productivity.

Challenges: Any new operating system must be robust and strong enough to tackle even the most challenging of problems. Moreover, it must be tested vigorously and upgraded routinely. Table 2 includes examples of such challenges, along with current progress.

Table 2. Challenges of the Enterprise Biology Software Project.

Challenges	Progress
Establish a production database that organizes, standardizes, and digitizes the data of the biology literature.	• Stereology Literature Database
Integrate data of the biology literature across disciplines – both physically and mathematically.	• Universal Biology Database • Hybrid Hierarchy Equation
Minimize bias and animal variation associated with published data.	• Universal Biology Database
Assemble a quantitative biology from the biology literature.	• Information Infrastructure
Build and test a digital model for biological research.	• Information Infrastructure • Yearly Report and Worldwide Distribution of the Enterprise Biology Software Package
Define guidelines for a quantitative biology.	• A Rule Book for Quantitative Biology
Forward and reverse engineer biology.	• Digital Libraries
Detect changes *in vivo* unambiguously.	• Digital Libraries • Experiments as Equations
Design and test a model for systems biology based on quantitative phenotypes.	• Digital Libraries • Systems Biology Two
Identify and crack organism codes	• Digital Libraries
Identify rules and principles of biology.	• See Table 3.

Rules and Principles: In biology, as in physics and chemistry, equations capture rules and principles explicitly (Table 3). In an information infrastructure, we need them to create order, to detect change, to assemble phenotypes, and to manage complexity. Without them, reinventing the biology literature becomes a largely impossible task.

Table 3. Summary of equations being used to estimate and interpret biological data.

Equations	Rules and Principles
Repertoire Equations	When taken two at a time, biological parts display distinct patterns of stoichiometry including valences.
Decimal Repertoire Equations	Repertoire equations can be fitted to decimal steps using regression analysis. They facilitate pattern recognition.
Design Code Equations	Parts in biology can change together as a group, even when individual rates differ.
Ladder Equations	Data in biology (CO+EX) be fitted to the linear portion of two intersecting growth curves, suggesting optimization at the core of biology.
Hierarchy equations	Biological data connect mathematically within and across hierarchical levels of size; a key feature of biological stereology.
Equations of the Experiment	Experimental design defined by balanced equations; variables become questions, evaluations answers.
Hybrid Hierarchy Equations	Relationships of structure to function can be defined mathematically in the design, execution, and interpretation of an experiment.
Engineering Equations	Unfolding and refolding complexity becomes the equivalent of diagnosis and prediction.
Phenotype Equations	A growing family of equations can be used to quantify phenotypes *in vivo*: simultaneous, regression, linear, power, polynomial, summation.
Connection Phenotype Equations	Local and global data sets expressed as polynomials.
Organism Code Equations	Rules and principles of biology encrypted in the parts and in their relationships. They provide networks capable of connecting all biological parts - quantitatively. See also Table 1.

Options

One of the central goals of this project is to take a hard look at the biology enterprise and to consider our options. Such a goal requires a willingness to ask tough questions and to follow through by pursuing credible answers. Everyone knows, for example, that our approaches to research in biology reflect largely what we want them to be.

Currently, the biology literature suggests that most investigators want a largely descriptive and semi-quantitative science. The implications of this choice become strikingly apparent when one assembles an information infrastructure for the biology literature. At every level of the experimental process, one quickly discovers numerous examples of investigators unwittingly minimizing the effectiveness of their research. In effect, we have trapped ourselves in a reductionist mindset that prevents us from addressing the central issue of our time, namely that of biological complexity. Figure 37 offers examples of our options. Option 1 suggests where we are today, whereas option 2 suggests where we can be. Since both options now exist in the real world, the choice becomes ours to make.

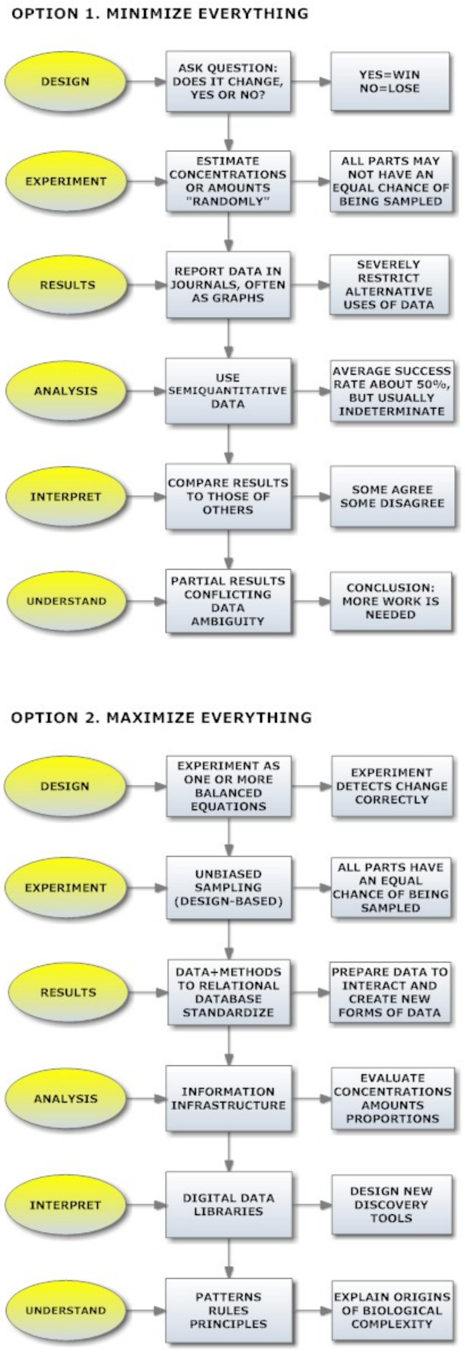

OPTION 1. MINIMIZE EVERYTHING
DESIGN
ASK QUESTION: DOES IT CHANGE, YES OR NO?
YES=WIN NO=LOSE
EXPERIMENT
ESTIMATE CONCENTRATIONS OR AMOUNTS "RANDOMLY"
ALL PARTS MAY NOT HAVE AN EQUAL CHANCE OF BEING SAMPLED
RESULTS
REPORT DATA IN JOURNALS, OFTEN AS GRAPHS
SEVERELY RESTRICT ALTERNATIVE USES OF DATA
ANALYSIS
USE SEMIQUANTITATIVE DATA
AVERAGE SUCCESS RATE ABOUT 50%, BUT USUALLY INDETERMINATE
INTERPRET
COMPARE RESULTS TO THOSE OF OTHERS
SOME AGREE SOME DISAGREE
UNDERSTAND
PARTIAL RESULTS CONFLICTING DATA AMBIGUITY
CONCLUSION: MORE WORK IS NEEDED
OPTION 2. MAXIMIZE EVERYTHING
DESIGN
EXPERIMENT AS ONE OR MORE BALANCED EQUATIONS
EXPERIMENT DETECTS CHANGE CORRECTLY
EXPERIMENT
UNBIASED SAMPLING (DESIGN-BASED)
ALL PARTS HAVE AN EQUAL CHANCE OF BEING SAMPLED
RESULTS
DATA+METHODS TO RELATIONAL DATABASE STANDARDIZE
PREPARE DATA TO INTERACT AND CREATE NEW FORMS OF DATA
ANALYSIS
INFORMATION INFRASTRUCTURE
EVALUATE CONCENTRATIONS AMOUNTS PROPORTIONS
INTERPRET
DIGITAL DATA LIBRARIES
DESIGN NEW DISCOVERY TOOLS
UNDERSTAND
PATTERNS RULES PRINCIPLES
EXPLAIN ORIGINS OF BIOLOGICAL COMPLEXITY

Figure 37 Current options in research biology.

Recommendations

Rationale: If we agree that mathematics and technology allow us to deal effectively with biological complexity and that such a capability is mission critical, then we can also agree on the need for major structural changes in the way we run our research enterprise. In truth, the study of biology is not about individual, isolated parts but rather about the relationships of all the many parts that contribute to defining the properties of an organism. Biology - by its very nature - becomes a deliberate study of complexity. As a research biologist, one can never be very far from the understanding that complexity is all about parts, connections, and relationships – and how they change relentlessly in health and disease.

If we expect the biology enterprise to deliver on its promises and responsibilities, then we need to upgrade our methods and practices to accommodate complexity. This will include a very careful look at the way we collect and report our data and how we plan to share them with our colleagues. Over the course of ten years, the Enterprise Biology Software Project has attempted to address these enterprise–wide questions and to provide helpful and insightful answers. The result of this effort has evolved into an information infrastructure for research biology, one that deliberately optimizes outcomes (Option 2., Figure 37). The effectiveness of this infrastructure rests firmly on a structural foundation provided by the mathematics of stereology and on the data produced by a scientific community well trained in the methods of unbiased sampling and in interpreting complex biological changes.

Recommendations: Establishing and maintaining a comprehensive and productive information infrastructure for the entire biology community requires a large-scale and long-term commitment. It becomes a strategy for success by protecting and increasing the effectiveness of our enormous investments in biomedical research worldwide. The following list of recommendations support this goal.

1. Report all research data of refereed publications in digital form as part of the publication process.
2. Institute a ten-year program with the goal of digitizing the biology literature by creating software for data entry and then making it freely available to investigators. Grants made readily available – especially to senior scientists - will actively encourage and strengthen this activity.

3. Use these digitized data to create a Universal Information Infrastructure, one capable of receiving data from all biological disciplines.
4. Use the biological hierarchy of size to design the structural core of the infrastructure and populate it with the most reliable data available. Currently, this will include largely stereological data collected with unbiased sampling methods.
5. Create resources that encourage investigators to use the information infrastructure to create software for their research program (e.g., digital libraries, applications), which can then be shared with colleagues.
6. Provide access to all the databases of the information infrastructure online, allowing them to be down loaded. Creating new digital libraries – basic to the discovery process - requires ready access to the database tables.
7. Establish funding support for finding a solution to the problem of counting molecules directly and quantitatively *in vivo*. Such a capability will be required to create a robust connection between the genome and the phenome.
8. Establish training programs in complexity for biologists. Creating digital libraries, for example, requires a working knowledge of data types, relational databases, and programming. Moreover, biological complexity defined in terms of parts connected quantitatively will require a solid foundation in biological stereology – both in theory and in practice.
9. Fund baseline studies for individual organs and organ systems designed specifically to collect and link data within and across hierarchical levels. Such resources, in turn, will support diagnosis and prediction in both the basic and clinical sciences.

Implications: Mining biology as a primary source of new technologies, information, and products will become a major source of productivity and success going forward. Indeed, the way we choose to manage our research and clinical data will define future health care, cost, and competitiveness worldwide. As a broadly based enterprise, we cannot expect success in solving problems of enormous complexity using methods and tools incompatible with complexity. The findings and accomplishments of the Enterprise Biology Software Project can and should be used by individuals and governmental agencies to argue that the construction of an information infrastructure and a quantitative biology is not only possible, but also mission critical to all segments of the biology enterprise.

Concluding Comments

Success in biology as an academic and commercial enterprise depends ultimately on our ability to deal with complexity. Although seemingly counterintuitive, the easiest and most direct way of reaching this goal is to allow biology to become a quantitative science, one driven by rules and principles. Biology comes to us packaged as two inextricable parts, a genome and a phenome. The genome contains the instructions for the phenome and the phenome includes all the phenotypes expressed by biological parts, including molecules, organelles, cells, tissues, organs, and organisms. Although it was possible to propose and fund a Human Genome Project without first solving the problem of biological complexity, the Human Phenome Project (Freimer and Sabatti, 2003) cannot go forward without a robust solution because the phenotype is the embodiment of complexity. With surprisingly few exceptions, health care, doctor-patient interactions, diagnosis, prediction, and our daily lives are all about phenotypes – not genotypes.

The most astonishing finding of this project is the speed at which complexity becomes simplicity when viewed through the lens of a quantitative biology. Since this robust form of biology appears automatically as the result of building an information infrastructure for the biology literature, what could be simpler or more appealing? Even the requirements for data entry become uncomplicated. A robust infrastructure welcomes all types of research data when they fulfill three minimum requirements. They must be collected with unbiased sampling methods, identified as variables and outcomes of the experimental equations, and related to a biological part.

Although rapid progress in any enterprise invariably requires bold steps, the rewards of innovation can be enormous. Consider, for example, the ever-present wonder of emergent properties. Once we have an information infrastructure for the biology literature, we also have a workable blueprint for the Human Phenome Project. The phenome becomes mission critical because we need it to understand the genome and vice versa. One without the other will never work effectively because it takes both to work out biological complexity. Moreover, the entire enterprise wins the moment we begin to play the complexity game.

REFERENCES

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

Bolender, R. P. 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 Enterprise Biology Software VIII. Research (2007) In: Enterprise Biology Software, Version 7.0 © 2007 Robert P. Bolender

Bolender, R. P. 2008 Enterprise Biology Software IX. Research (2008) In: Enterprise Biology Software, Version 8.0 © 2008 Robert P. Bolender

Bolender, R. P. 2009 Enterprise Biology Software X. Research (2009) In: Enterprise Biology Software, Version 9.0 © 2009 Robert P. Bolender

Douady, S. and Y. Couder 1996 Phyllotaxis as a Dynamical Self Organizing Process (Part I, II, III), J. Theor. Biol. 139, 178-312.

Freimer, N. and C. Sabatti. 2003 The Human Phenome Project., Nature Genetics 34, 15 – 21.

Günesdogan, U., Jäckle, H., and A. Herzig. 2010 A genetic system to assess in vivo the functions of histones and histone modifications in higher eukaryotes., EMBO reports 11, 772 – 776.

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