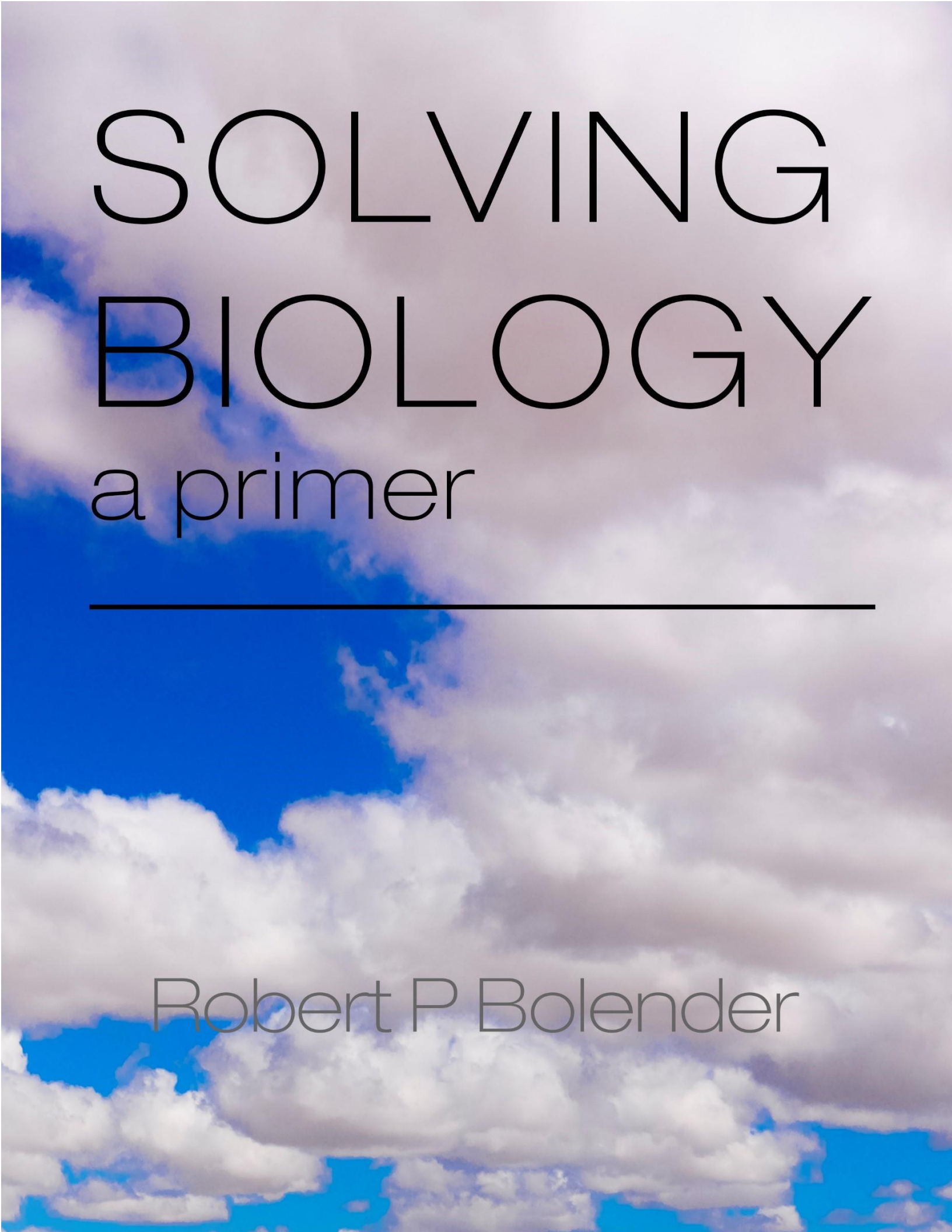

SOLVING
BIOLOGY
a primer
Robert P Bolender

SOLVING BIOLOGY

a primer

Robert P. Bolender

Printed in the United States of America

First Printing, 2019

ISBN 978-0-9971543-3-7

Enterprise Biology Software Project
PO Box 166
Medina, WA 98039

www.enterprisebiology.com

www.playingcomplexitygames.com

www.solvingbiology.com

PREFACE

Is it possible to solve biology? Yes, but the answer we want involves rethinking the way we approach biology as a science.

As a product of nature, biology adheres to a theory structure governed by well-established rules. Such rules – unknown to us for the most part - have a direct bearing on how we collect, express, and interpret our data. Of interest to us here is to figure out how to play biology's game according to biology's rules.

THE PACKAGE

Solving biology begins by addressing the complexity problem. To do this, we will need access to biology's package of skills along with directions explaining how to find, interpret, and use them. The package includes a skill set designed to manage stability and change within the context of precision, accuracy, diagnosis, prediction, and reproducibility.

The plan is as follows. We'll collect reprints from the biology literature, transfer the data to a relational database, recover the missing connections and complexities, generate universal biology databases, use these databases as parallel complexities to copy biology's rules, apply the rules to access the package, and then begin the process of solving biology. The primer explains the process and serves as a guide for the beginner.

THE OPTIONS

Currently, we have two research options: Plans A and B.

PLAN A: Biology, chemistry, and physics all share the same theory structure, called reductionism. It simplifies a complexity by taking it apart so that we can study the properties of the individual parts. If we understand the parts, the theory argues, we can understand the complexity because the whole is equal to the sum of the parts.

When applied to a complexity such as biology, however, this theory is not a good fit with reality because the whole turns out to be greater than the sum of its parts. By reducing biology to just a collection of parts, we deliberately discard two of biology's key properties - connectivity and complexity. This amounts to throwing away more than 50% of what biology has to offer us. In return for the many conveniences provided by reductionism, we pay a steep price. The remaining data, which characterize the parts, tend to be noisy, biased, and incomplete – hardly up to the demanding task of solving biology.

One of the incentives for writing this primer was to come up with a workable plan B.

PLAN B: The alternative plan starts with a wish list determined to leave the limitations of Plan A far behind. Instead, we want biological data that are quiet (displaying little or no variation), largely unbiased, and nearly complete. Moreover, we also want unfettered access to biology's package. To go from Plan A to B, we will have to figure out how to make the transition from a theory structure based on reductionism to one based on complexity.

Accordingly, we can proceed by setting out two primary objectives for the primer. It must first explain how to give biology back its missing connections and complexity and then make a

convincing argument for the plausibility of solving biology mathematically. This will require showing that taking biology apart produces one result, whereas putting it back together produces a different but much better result.

TOWARD A COMPLEXITY THEORY: If progress in biology depends on coming up with a more realistic theory structure, then how do we define it? Fortunately, we don't have to because it already exists. We're just going to figure out how to copy the one biology uses. Since we know *a priori* that complexity theory works for biology, it follows that it should work for us as well.

By copying instead of inventing, we will come face to face with a curious paradox. When playing biology's game, it turns out that complexity makes hard things easier, whereas reductionism makes hard things considerably harder. Why? By requiring assumptions to compensate for the data it discards, reductionism has a destabilizing effect. Indeed, it bears a large responsibility for the current disorderly state of the biology literature.

Given the facts on the ground, simplifying complexity by throwing away vital information no longer makes much sense. Why? We now have the technology and the wherewithal to move up to complexity where all the key elements can be put in play.

STARTING FROM FIRST PRINCIPLES: We will start with some of the most complex structures of biology, disassemble them into their basic parts, gather their underlying rules, and then reassemble them to discover how biology is built on a mathematical foundation. We will use these rules to make discoveries, identify first principles, and begin to assemble a new theory structure.

The biology literature comes first. We'll reduce it to a table of parts (identified by names and numerical values), reconnect the parts by forming ratios, and then use new data types to redefine the literature, to diagnose clinical disorders objectively, and to see how biology changes and generates complexity. Such training exercises will hone our skills at complex problem solving.

Next, we will attempt something far more difficult – explain a biological change from first principles. We'll begin this task by first using data coming from individual animals to show how we can use biological variation to unmask biology's reproducibility rules and then continue by encouraging biology to explain – with rules – how it uses relationships of structure to function to bring about complex changes.

Notice where this is leading. By knowing how to explain a biological change from first principles, we can work toward the goal of characterizing phenotypes with rule-based equations. Such phenotypes become helpful because they can ease the much harder task of reverse engineering gene expression.

HOW TO USE THIS BOOK

THE SCOPE: This book picks up where an earlier one left off (Bolender, 2016b). The first book took the story only as far as the first level of complexity, whereas the primer adds five new levels. This takes us into the meaty part of complexity where we get to engage problems fundamental to the theory structure of biology. These include substantial topics, including reproducibility, the postulates of biochemical homogeneity, clinical diagnosis, prediction, change, and the inextricable relationship of structure to function. By showing that mathematical solutions can exist for such weighty topics, we will advance toward our goal of solving biology.

THE BIG PICTURE: The book addresses each of the substantial topics listed above as they relate to specific levels of biological complexity. In effect, the chapters become training sets designed to assist the beginner in seeing how biology operates by embedding one complexity in another. Taken together, they supply a preliminary view of the big picture.

Since the primer targets beginners, it cannot be an advanced textbook. However, extensive details exist elsewhere (enterprisebiology.com, playingcomplexitygames.com). These online resources include yearly reports (2001-2018), databases, spreadsheets, worked examples, and step-by-step instructions.

THE STRATEGY: The primer focuses its attention primarily on the way biology does things mathematically. The premise underlying the Enterprise Biology Software Project, of which this book is now a part, is that biology already knows the answers to most of our questions. Consequently, this reduces our job to one of figuring out how to ask questions that biology can answer in ways that we can understand and, in turn, use to our advantage.

How does this work? Biology responds to our questions (submitted herein as literature-based experiments) by applying rules - the results of which we will be capturing and interpreting with images, patterns, and equations. In effect, solving biology becomes an exercise in upgrading the biology literature in ways that allow us to tap into the biology's clever approaches to problem solving. For us, the rules will become equations, which in the context of a theory structure, play a key role in defining the properties and the standing of a science.

THE BENEFITS: Copying biology as a problem-solving strategy offers a surprisingly straight forward approach to complexity. In turn, complexity offers a wealth of additional benefits. We will use it, for example, to explain why many of the problems currently plaguing biology are of our own making. By choosing a quantitative approach to biology instead of the traditional descriptive one, our job of solving problems becomes much easier, more defensible, and far more rewarding.

THE CATCH 22: By operating largely under the rules of chemistry, cell and molecular biology face an untold number of problems when trying to interpret biological data. During a change, for example, knowing what's happening chemically to biological parts does not necessarily translate into knowing what's happening biologically. The problem is one of simple versus complex. While chemistry focuses on just parts, biology in real life focuses on parts, connections, complexity, and emergent properties. As a result, chemistry and biology often tell very different stories for the same experiments.

Nonetheless, we are still expected to take biology apart and then use the resulting data to interpret complex changes that can occur only in an intact, living organism. Such an impossible requirement creates a catch 22. Our best chance of avoiding this catch 22 is to figure out how to put biology back together.

THE THEORY STRUCTURE: All the sciences operate according to a theory structure, one that includes a collection of guiding principles and procedures. To do biology, however, we need at least two theory structures - one to connect biology to physics and chemistry (reductionist theory), and another one to copy and solve biology mathematically (complexity theory).

Why? Physics and chemistry operate according to rules instigated by the elements of the periodic table, whereas biology must accommodate rules for both the periodic table and a massive table of genes responsible for producing and operating phenotypes.

Consequently, complexity theory is needed to merge the properties of both tables and to grasp the relationship of one table to the other.

In short, the theory structure envisioned herein for biology derives from a data-driven approach wherein permission and affirmation will – by necessity and convention - come from the basic principles defined by nature.

ACKNOWLEDGMENTS[1]

"The decision to approach biology as a complexity came from a month-long workshop held in Santa Fe, NM (1987) under the auspices of the Santa Fe Institute and funded by the NIH, NSF, and Simon Foundation. It was held in response to a recommendation of the National Research Council (1985). Our group was charged with the task of figuring out how to organize all the published data of biology in such a way as to reveal generalizations, connections, and new theory structures. The effort produced a strategic plan accompanied by a list of recommendations (Morowitz and Smith, 1987).

In turn, the insights, guidance, and enthusiasm provided by this workshop led to a pilot study (Bolender and Bluhm, 1992) and then to a grant from the National Science Foundation (NSF). The goal of the NSF grant was to organize the published data of biological stereology within the framework of a relational database. This grant along with helpful suggestions from the NSF provided the foundation for the ongoing Enterprise Biology Software Project (2001-present), which includes yearly progress reports, et cetera. The mission of this project is simple, but critical: to provide rigorous and objective support to students entering research disciplines in the biomedical sciences.

The success of this project can be attributed to the generosity of the stereology community for supporting the project and supplying reprints for the stereology literature database and to the Internet Brain Volume Database (Kennedy, et al., 2012) for providing online access to the MRI data of patients. Since many of the secrets to understanding biology as a complexity exist in the biology literature, we will use the primer to explain how to translate these secrets into a wide range of solutions.

[1] Reprinted in part from Playing Complexity Games (Bolender, 2016b).

INTRODUCTION

A primer is a little book of instruction designed to introduce the beginner to a new subject. In our case, we will use it to explain how interacting with biology as a complexity is akin to solving it. The process is well-defined. Assemble a literature database, reformat the data into parallel complexities, capture biology's rules with equations, and let the equations solve the problems.

Chapter one sets the stage by telling the reader what to expect. By structuring the book around levels of complexity, the narrative can focus on data, databases, software, and mathematical arguments – not as a collection of separate players, but instead as a team charged with the tasks of reconnecting parts, recovering lost information, finding secrets, and delivering solutions. Our strategy becomes the interesting part. When we give biology a problem to solve, we will copy whatever biology does in response by generating parallel complexities.

A basic theme pervades the book. We cast biology as a rule-based complexity, one that uses parts and connections to define relationships of structure to function. In a biological complexity, wherein everything is connected mathematically, help will always be available wherever and whenever we need it.

By embracing this premise, we get to use biological complexity to our advantage in ways that begin to redefine our understanding of biology as a science. This puts us on a collision course with the single, biggest intellectual challenge biology faces as a science – the inevitable shift in our thinking and action from descriptive to objective. The primer will argue that the one thing needed to make this happen is a willingness on our part to interact with biology as a complexity. If you – as a beginner – are willing to give it a try, then you will no longer be a beginner.

Although biological complexity sounds daunting, it turns out to be quite the opposite. Biology offers unflagging encouragement by operating with a clarity based on rules that it expresses mathematically. The whole point of mathematics is to make difficult things easier and understandable. As a complexity, biology represents the ideal candidate to become a quantitative science because it does things by rule. Since we can get our hands on these rules by fitting published data to equations, we will be interacting directly with the core principles of biology. In effect, we will be deriving biology from first principles – following the same strategy used so successfully by physics and chemistry.

STRATEGIES

Two linear sequences help to summarize the story. They both start with the literature and lead either to principles or to products.

PRINCIPLES

Biology literature → organization → integration → patterns → equations → rules → principles.

PRODUCTS

Biology literature → organization → integration → products (patterns, equations, rules, change, diagnosis, prediction, and reproducibility).

Both sequences describe pathways into biological complexity that lead to local and global solutions, including verifications.

THE PACKAGE

The challenge of any science is to be in control of the six items listed in the package shown below (Figure I.1). Although biology is the most complex of the sciences, it lacks a strategy for managing complexity and struggles to make the grade with all six skills. This puts biology on shaky ground because such skills are the bread and butter of a quantitative science. Accordingly, our task of solving biology becomes one of sliding our current descriptive platform into a quantitative one.

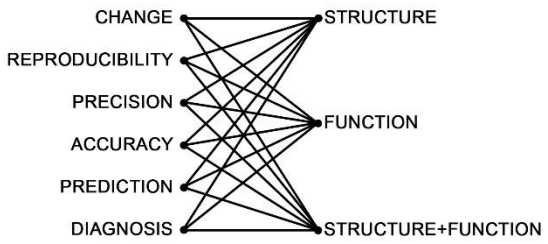


Figure I.1 The package. As a science, biology depends importantly on providing investigators with six critical skills, which include detecting biological changes, reproducing results, assuring the precision and accuracy of data, and predicting and diagnosing events. Understandably, an exercise in solving biology begins by probing the relationship of our structural and functional data to these skills.

It is incumbent on you – the beginner - to become aware of the shortcomings of your science and to understand that the job of setting things right invariably belongs to each new generation. Such is the nature of progress.

The primer offers the beginner a head start. In biology, we tend to specialize - some investigators prefer structural approaches (morphology) others functional (biochemistry). Consequently, it will be useful for you to know how well these two disciplines perform vis à vis our list of six essentials (Figure I.1).

Remember that acquiring the package is the prize. Biology already has it, we want it, but it remains stubbornly beyond our reach. The primer explains why this is the case, promptly admits defeat, and defaults to the simplicity of an approach based on copying the package from biology. We will do this by taking a page out of the physicist's playbook. When encountering an impasse, they come up with a new theory structure or tweak an existing one. We'll play the same game, except that we'll leave it up to biology to change the rules and supply us with a new theory structure. Since biology already possesses all the skills listed in the package, we will argue that it is the most qualified to become our mentor and role model.

LEVELS OF COMPLEXITY

Although we describe the biological hierarchy as a collection of parts, ranging in size from molecules at one end to organisms at the other, biology prefers to arrange its operation hierarchically as continuous sets of complexities embedded in complexities. This works to our advantage because solutions (expressed as patterns or equations) will exist everywhere we look.

By copying biology, we will be able to upgrade our traditional biological hierarchy of parts to one of complexity by bundling both parts and connections within the same quantitative framework. Such an approach sets the stage for generating outcomes based on generalizations, rules, and first principles.

Thus far, we can identify six levels of complexity populated with specific data types, equations, and solutions. Level 1 involves quantitative patterns relating structure to structure or

function to function, whereas levels 2 to 6 focus on relationships of structure to function.

What do these levels contribute? When we want to detect and interpret a biological change such levels unfold a complexity into well-defined sequences of interpretable events. The story of how a biological change occurs is a compelling one because it lets us open and look inside the black boxes that continue to be largely responsible for generating our experimental results. Instead, we will interpret a biological change as it cascades from one level of complexity to the next by collecting the rules as they come into play. Such an exercise is helpful because we can use it to show how quickly discovery leads to innovative solutions.

REDISCOVERING BIOLOGY

We are beginning to understand that discovery in the life sciences depends increasingly on our having ready access to substantial amounts of published data that we can reformat and filter in ways designed to target and deliver robust solutions with staying power.

Since biology organizes itself structurally, we will begin the discovery process by upgrading the literature of biological stereology within the framework of a relational database. Recall that stereology is a method widely used to quantify structure. It includes a collection of design-based methods based on probability theory that estimate quantitative measures of biological parts (including, for example, volumes, surfaces, lengths, and numbers) from measurements made on tissue sections. The sampling methods, which are unbiased and wonderfully clever, allow us to make estimates from intact tissue samples comparable to the unbiased estimates coming from the tissue homogenates of biochemistry. Were this not the case, finding quantitative relationships of structure to function would not be possible.

Once designed and populated with published data, we can use the stereology literature database to begin the process of figuring out how to approach biology as a complexity. The first order of business is to assemble enough evidence to convince us that structural data adhere to an underlying mathematical order. A typical approach to such a task consists of fitting data points to equations with regression analysis, wherein the gate keeper becomes the coefficient of determination (the R^2 value).

We begin with the obvious question. Will data taken from different papers and species sit on the same regression line and display R^2s equal to or close to one? A yes answer is helpful, a no answer is not. Fortunately, the literature provides biological data that fit regression curves with R^2 equal to or close to one often enough to encourage us to push on.

The next question is a harder one to answer. How do we get complexity from a database populated with just parts? What do we do?

If biology exists as a complexity, then the stereology literature database might contain such complexities or at least harbor residues of their antecedents. First, however, we need to understand what constitutes a complexity.

The dictionary defines complexity as the relationship of parts to connections. Does this mean that we need two types of data – parts and connections - to study a complexity? Yes. Do such data exist in the biology literature? No.

This tells us what needs to be done next. We must figure out how to add the connections back to the parts and then show that the resulting complex patterns (parts + connections) replicate often enough within and across species to be convincing. In effect, we

need to show that reproducibility exists as a general property of the biology literature.

Fortunately, an army of regression equations bearing R^2s ≈ 1 and passing through or near the origin provided the essential clue to the connection problem. Any point on such a regression line defines the same numerical relationship of one part to another as a ratio (x:y). In effect, the numerical ratio defined by the two parts serves to connect them mathematically. This puts connectivity back in the game.

For example, a data pair (ax:by) consisting of two parts (a, b) and one connection (x:y) defines a complex data type. To increase the complexity of the data type, we simply add more parts and connections to the alphanumeric string.

By interpreting these new complex data types with programs designed specifically to find patterns and complex relationships, we will discover that the data types of level 1 complexity deliver troves of brand-new information. More importantly, they will encourage us to dig deeper.

Discovery and Innovation

As the story unfolds, we'll be moving from our current research platform with its well-established rules and methods to one where the rules, terms, and problem-solving strategies remain in flux. Although at first this may be disorienting and even daunting, any initial uneasiness will quickly pass.

As we move into the n-dimensional spaces inhabited by biology, things happen differently and in ways sometimes inconsistent with what we have been taught. Such a passage, however, allows us to enter biology's world where it lives and routinely solves problems that to us – looking from the outside - can appear unfathomably difficult. By taking a "down the rabbit hole" approach to discovery and innovation, we can escape the limitations of our current reality and step onto new platforms operating under theory structures already thoroughly vetted by biology. The following examples of this seemingly curious strategy will help.

Biology Literature

Our first job consists of opening the scientific literature. This requires changing its configuration from that of a holding cell for our research data to an interactive and productive research tool. Instead of having to track down data by laboriously browsing through piles of research papers and then translating the collected data into usable forms, we want all the data to share the same format and to be universally connectable and searchable. Assembling and using such universal biology databases becomes one of the first things you will learn to do.

Diagnosis

Since universal databases adhere to biology's rules of organization, we can use them to explore solutions to uncommonly difficult problems. Consider, for example, clinical diagnosis. Even the most knowledgeable of physicians routinely disagree with their colleagues on a given diagnosis because disorders display so many of the same symptoms.

This triggers the obvious question. Is it possible to upgrade clinical diagnosis? To answer this question, we must figure out how to extract the diagnostic skills of thousands of highly skilled physicians, quantify these skills, and store them

as diagnostic codes (alpha-numeric strings) in a database. In turn, we can use the database as a diagnostic tool to identify a disorder by solving the complexity puzzle presented to us by a patient. The hard part of this problem will consist of coming up with an acceptable solution to a central problem of clinical diagnosis – the glut of uncertainty produced by false positives and false negatives.

We will use this case study to introduce you to the topic of big data as it applies to clinical diagnosis. Moreover, you will get to see how biology constructs itself - by rule - in health and disease and how we can use such rule-based constructs as unique identifiers.

CHANGE

A biological change often appears chaotic because it defines a time when the rules are changing. The central challenge in explaining a change begins by developing the skills needed to capture the rules biology applies – when, where, and for how long. Having such a skill set, for example, will make it possible for us to reconstruct the quantitative phenotypes needed to explore and reverse-engineer gene expression.

One of the major challenges of 21st century biology will be to explain the relationship of the genome to its highly adaptable phenotypes.

Although change is somewhat of an advanced topic for a primer, it will be included because it provides a hands-on opportunity for you to try out your newly acquired package of biology-based skills.

BASIC ASSUMPTIONS

WORKING DEFINITIONS

Basic truths can serve an important function by defining the ground rules under which the primer operates. They are as follows:

- *Biology exists as a complexity.*
- *Organization is the first step to understanding complexity.*
- *Solving biology is an exercise in exploring complexity by creating complexity.*
- *Biology already knows much of what we want to know.*
- *All three sciences – physics, chemistry, and biology - can be derived from first principles.*
- *The one language shared equally by the natural and physical sciences and by those wishing to study them is mathematics.*
- *In the sciences, rules of evidence can include $R^2 = 1$, $R^2 \approx 1$, diagnosis, prediction, and reproducibility.*
- *What's true for the individual is true for the population.*

CHAPTER 1

LEVELS OF COMPLEXITY

Solving biology is an exercise in solving complexities by creating complexities. The process consists of using published data to create complexities parallel to the ones used by biology to solve a given problem. In effect, we recruit biology to solve our problems for us, being confident that it has already worked out the best solutions. We can safely assume that biology is qualified to do the heavy lifting because it knows what rules to apply and how to summon the necessary resources. It understands full well that survival depends on its ability to adapt quickly and effectively.

Chapter 1, which introduces the first six levels of complexity, tells the story simply and succinctly. Although touching on the central themes and principal findings, it keeps details to a minimum. Getting the big picture at the outset will make it easier for us to tackle the specifics in the ensuing chapters.

1.1 LEVEL 1 – PATTERNS

Level 1 complexity includes patterns based on either morphological or biochemical data, but not both data types together. Two databases – one populated with stereological data (Bolender, 2001a) and the other with clinical MRI data (Kennedy et al., 2012) supply the raw material for generating complex patterns.

Translating data sets from simple to complex begins by collecting all the connectable data of a given paper (namely data related to the same reference) and, in turn, using them to generate all possible permutations of the parts taken two, three, or four at a time. This generates a collection of complex data types identified as data pairs (ax:by), triplets (ax:by:cz), and quadruplets (ax:by:cz:dq). We call these alphanumeric strings mathematical markers. Think of them as snippets of phenotypic complexity taken from much longer strings and networks. Let's look at some examples of what they can do for us.

1.1.1 Biological Blueprint

By generating data pairs (ax:by) for all the biological parts stored in the stereology database, the resulting patterns show how biology constructs itself according to a modular design. The patterns, for example, created by the data pairs tend to generalize (i.e., many duplicates appear) within and across species, thereby suggesting that the basic design of animal phenotypes is widely conserved.

It soon becomes apparent that these linear strings made up of parts and connections can serve as mathematical markers that simplify our job of working out solutions to a wide range of problems. For example, they prove to be well-suited to the task of dealing with diagnosis and prediction in clinical settings. The mathematical markers supply not only objective results, but they can also deal effectively with the uncertainties produced by false positives and false negatives.

1.1.2 Disorders of the Brain

When based on subjective judgements, a clinical diagnosis often becomes challenging because closely related disorders tend to display similar symptoms. Consequently, physicians making a given diagnosis all too often fail to agree with their colleagues. A workable solution to such a dilemma consists of assembling an objective test based on a set of patterns generated from the clinical literature.

By translating the MRI data of the Internet Brain Volume Database (IBVD; Kennedy, et al., 2012) into mathematical markers, for example, large numbers of objective patterns become linked to the diagnostic findings of expert physicians. In turn, we can avoid the problem of false positives and false negatives by requiring that all the mathematical markers associated with a given disorder are unique to that disorder. In effect, the database becomes a parallel complexity capable of solving the challenging problem of clinical diagnosis.

How does such a diagnostic procedure work? The numerical construct consists of encapsulating the mathematical markers in a data cage, wherein selecting one or more test markers therefrom will always result in the correct diagnosis because it is the only outcome allowed by the design of the database. Such a strategy is deliberate. It serves to verify the effectiveness of thousands to millions of markers operating in a well-defined and closed setting.

Notice the strategy of the approach to problem solving. A diagnosis depends on the preponderance of the evidence. If we start with a diagnostic database containing millions of unique and verifiable markers, then the likelihood of diagnosing a patient correctly with markers outside the cage will be much improved. In effect, by using a big data approach, the data cage attempts to optimize the diagnostic process.

Consider the likely spinoffs of this objective approach. Evidence, for example, is beginning to accumulate that the quantitative organization of the brain can phenotype the state of an individual, analogous to the way a fingerprint provides the identity of an individual. Consider the implication of such an insight to health care.

1.1.3 Health Care

If, as the data cage example suggests, mathematical markers offer an effective diagnostic tool for disorders, then the brain might also be carrying patterns ideal for assessing our current and future health. In fact, reports of head scans being used to diagnose systemic disorders already exist in the literature (Cecil et al. 2008, Guido et al. 2013, Herting et al. 2014, Khan et al. 2011, Strassburger et al. 1997, and Tiehuis et al. 2008).

With the eventual introduction of portable MRI scanners linked to diagnostic databases, routine monitoring and feedback could become commonplace. Such a device would compare the patterns of an individual's brain to a library of patterns linked to expected behaviors and outcomes. This juxtaposition of the present to the past to the future provides an objective link between diagnosis and prediction and is likely to spark widespread innovation.

1.1.4 Gold Standards

A frequent – although questionable - practice of biomedical research is the willingness to mix results coming from living and nonliving sources. However, this creates a problem in that each data type and estimate carry unique sets of methodological biases. Stereology is no

exception in that it suffers a host of distortions that occur when specimens are prepared for microscopy (Weibel and Paumgartner, 1978, Bertram et al., 1986).

A solution to the living vs. nonliving data problem consists of generating correction factors, by allowing data derived from living sources to serve as gold standards. In effect, this allows data derived from living and nonliving sources to give comparable results. Harmonizing data sets makes good sense because it sharpens patterns and creates bridges spanning methods, publications, and levels of complexity.

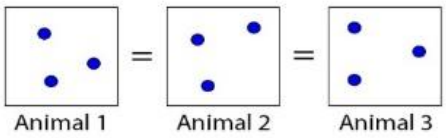
1.1.5 Connectivity Patterns

Since biology uses ratios to order its parts, we can expect to find patterns of similarity within and across species - throughout the literature. Moreover, patterns detected with mathematical markers provide fresh insights into how biology organizes its parts and adapts to change. With molecular biology continually reminding us of our shared genetic roots, finding persistent patterns among phenotypic parts provides a welcome reassurance. Using methods of cluster analysis, for example, we will discover that the biology literature contains large numbers of highly conserved patterns.

1.1.6 First Principles

When the same pattern repeats across varied species and settings, it signals the presence of an underlying first principle. Mathematical markers, which were designed to detect biological patterns quantitatively, tap into biology's ratios and uses them to define a basic unit of biological complexity. Such ratios, which define the relationships of parts to connections, identify a first principle of biology. In fact, biology derives much of its complexity from ratios.

1.2 Level 2 – Biochemical Homogeneity



Equations play a key role in biological complexity by identifying relationships among biological variables and constants. For our purposes here, finding an equation locally (it exists in one publication) predicts a global pattern (it exists in many publications). This means that rule-based equations provide a built-in test for reproducibility.

1.2.1 Postulates of Biochemical Homogeneity

Our understanding of biological cells and the methods we use to study them relies importantly on the postulates of biochemical homogeneity, as put forward by Christen de Duve (1964, 1974) and his colleagues. These postulates state that biochemical activity distributes uniformly at a single morphological location. Recall that a postulate - akin to an axiom – accepts something as true without theoretical or empirical proof.

Our tacit acceptance of the postulates defining the relationship of structure to function in cells, however, identifies a new problem for us to solve. In effect, we need to figure out how to translate published data into equations that capture and test these postulates empirically.

Once demonstrated, the postulates can become first principles.

1.2.2 Mathematical Relationships of Structure to Function

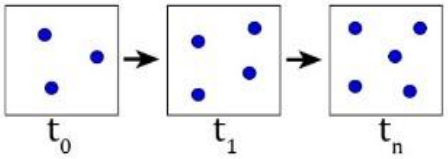
The brilliance behind de Duve's postulates was that they established a quantitative link – albeit it a putative one - between a biochemical measure (enzyme activity) and a specific morphological location (cell organelle). Although the postulates contained the key elements needed to connect the phenotypic parts of an organism to their parent DNA, it failed to happen because the technology of that earlier time was not up to the task of collecting evidence from large data sets. Instead, the data of biochemistry and morphology became widely correlated, but not connected to one another by rule.

Why not? We didn't know where to look. If, for example, we plot the surface area of the ER in hepatocytes (estimated stereologically and related to a gram of liver) against the biochemical activity of an ER marker enzyme (also related to a gram of liver), we get a linear regression with a very disappointing R^2 value. Few, if any, of the data points sit on the regression line. Since such an outcome is typical of structure-function plots, the very existence of such a relationship remained tentative. There was, of course, an alternative explanation. What if biology doesn't do relationships of structure to function the way we do them?

If, instead, we recall that biology likes to organize itself with ratios, we can follow this clue, adhere to the postulates of biochemical homogeneity, and rerun the regression. Now we can get an equation with an $R^2 = 1$ (all the data points are sitting on the regression line).

Although such an equation gives the expected result, it still needs verification. When tested, however, it passes both the reproducibility and prediction tests at the global level. This tells us that we can capture relationships of structure to function with equations capable of copying biology's rules. In short, level 2 complexity explains how biology expresses relationships of structure to function mathematically – by rule.

1.3 Level 3 – Organelle Changes



We – all of us – have been taught how to compare two states of a part – control and experimental – to see if the part changed. For biology, however, changing a part is a much bigger operation involving individual changes in parts and connections at multiple levels of complexity. If we want to view a change through biology's eyes, then we need to copy what biology does at all the levels of complexity in play. In effect, complexity levels 3 to 6 attempt to explain – at least as a first approximation – how biology engineers a change.

The understanding to take from the complexity levels exercise is that detecting a change in a biological part the usual way by quantifying the behavior of a selected part does little to explain what in fact happens. Although the widely accepted approach to detecting biological changes involves black boxes and assumptions, the alternative strategy we will be using here relies instead on equations, transparency, and reproducibility.

1.3.1 Steady State (Control)

To understand how a biological change occurs, we will look at the rules in play during three states: steady state (no change), transitional state (changing), and new steady state (change completed). Let's tackle the hard one first – the steady state in the control. Although the equations of Level 2 can demonstrate that quantitative relationships of structure to function exist, we have yet to present enough empirical evidence in support of de Duve's postulates (i.e., that membrane bound marker enzymes distribute uniformly at a single morphological location).

For the postulate to be correct, we would at least have to show that the marker enzyme activities distribute equally across both the rough and smooth-surfaced components of the ER. In other words, the ratio of the enzyme activity associated with a unit of SER surface area (1 m^2) should be the same as the enzyme activity associated with a similar unit of RER surface area.

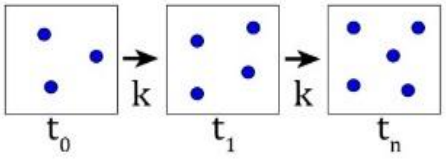
1.3.2 The Biochemical Homogeneity Test

To run the homogeneity test, we need to write simultaneous equations that solve for the concentrations of the enzymes in the smooth (SER) and rough (RER) membranes of the endoplasmic reticulum (ER). However, the results of this test showed that the SER membranes had a 7% higher concentration of the marker enzymes than those of the RER (P=0.0056).

Does this finding of a heterogeneous distribution of a marker enzyme on the two ER subcompartments disprove the uniform distribution (homogeneity) postulate? No, not at all. The RER loses membrane surface area to ribosomes that are bound to it, but not to the SER. It seems more likely that the 7% difference between the two enzyme densities estimates can be explained by the amount of RER membrane allocated to the ribosomal attachment sites. In short, the principles of biochemical homogeneity appear to be on solid ground – empirically.

Notice that level 3 complexity (change) depends on the underlying complexity of level 2 (biochemical homogeneity). Every time we fold in a new level of complexity, new opportunities arises. Biology already knows this and makes effective use of it. To have connectivity across complexity levels, the order of a given level must extend to both the previous and subsequent levels. This folding of complexity – one level embedded within another – explains how complexity changes as it cascades up and down biology's hierarchy of size. It follows that such a Gordian knot should be allowed to untie itself.

1.4 Level 4 – Rates of Change



Since biological cells derive their properties by repeatedly embedding complexities within complexities, a change changes many things. Not the least of which, of course, are the rules in play. By introducing the variable of time into the mix, cells unleash a host of cascading variables wherein seemingly endless arrays of changes are occurring within changes. Complexity becomes rampant. That said, it follows that capturing the rules as they appear during a biological change becomes an important contributing factor to the success of an experiment.

1.4.1 Transitional State (Change)

After triggering a change, the cell turns on a complex array of machinery that quickly redefines the structural and functional properties of its parts and connections. Once the cell adapts to its new reality, the transitional state ends, and a new set of steady state rules appears. In effect, changing the state changes the rules.

Figure 1.1 summarizes the process. It gains support from the observation that for a given set of parts different mathematical markers characterize the control and experimental states.

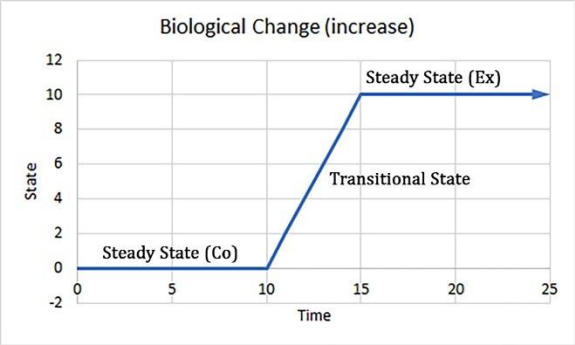
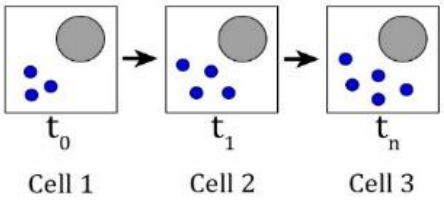


Figure 1.1 As a working hypothesis, we can define a biological change as two steady states separated by a transitional state.

A workable approach to the change problem begins by treating it as two mathematical events occurring simultaneously, both of which can be defined by equations. The first set of equations (event 1) treats each experimental time point as we did for the control data of Level 2 Equations, whereas the second set (event 2) focuses on the rate of change identified in levels 3 and 4. In effect, this reduces the process of tracking a change to looking for linear regressions with R^2s equal to or close to one. As expected, level 4 complexity depends on the complexities found in levels 2 and 3. This is in keeping with the complexity embedded in complexity model.

The equations, for example, will tell us that the concentrations of the marker enzymes in the ER membranes of hepatocytes change over time (event 1) and that the rate of change (event 2) is consistent with zeroth order kinetics. Since the regression analysis delivers rate constant equations with R^2 equal to or close to one, we are effectively copying biology's rules as they appear and orchestrate the ongoing change.

1.5 Level 5 – Cell Changes



In the next two levels, the complexity shifts from detecting changes in average cells (level 5) to detecting changes in the organs containing such cells (level 6). Using stereology, we can estimate average cell data by dividing the total cell volume by the number of cells contained therein. Biochemistry does not have a similar average cell option and consequently must default to making assumptions.

Level 5 complexity is the point at which many research papers run into trouble because detecting average cell changes *in situ* and maintaining reliable data references are capabilities not available to many of our current research methods. This limitation comes into sharp focus once we set out to solve biology.

The primer describes two attempts to address this problem objectively with equations. The first equation takes a slightly convoluted approach, whereas the second one leverages a biological rule found at level 2 complexity.

Nonetheless, they both can produce the same result.

1.5.1 Detecting Average Cell Changes

We can manage problems encountered at level 5 complexity by exploiting a connection that exists between changes related to average cells and to the organ containing these cells. If the number of cells in the organ remains constant, both estimates (per average cell and per total organ volume) will detect the same relative amount of change.

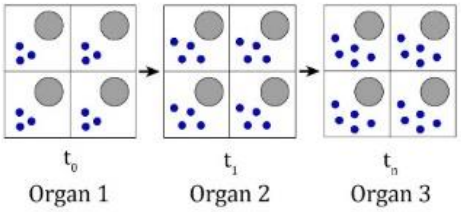
1.5.2 Detecting Cell Changes with Concentrations

If, during an experiment, the total number of cells of interest remains constant (little or no mitosis or cell death occurs), but the volume of the average cell changes, what happens to the cellular content of a cubic centimeter or gram of a tissue or organ? If the volume of the average cell decreases, it takes more cells to fill a cm^3 of reference space, but, conversely, fewer cells will be needed if the average cell volume increases. When comparing concentration data, both events introduce a bias, which we must take into consideration.

Consider a simplified example. If the volume of the liver increases by 10% all of which can be attributed to an increase in the volume of an average hepatocyte, then to detect a change using concentrations, we would still have to compare data coming from the same number of hepatocytes. Since a 10% increase in the volume of an average hepatocyte pushes 10% of the hepatocytes out of a cm^3 of hepatocytes, increasing the value of the concentration by 10% theoretically pushes the cells back in. We can do this because the following statement is true. Comparing changes in concentrations is equivalent to comparing changes in average cells when the number of cells being compared remains constant.

The advantage of this approach is that the same equation works for both morphological and biochemical data by providing the equivalent of average cell data without having to count cells. The equation, however, requires assumptions to be discussed later. The more compelling equation would be one that avoids having to make assumptions. Why? Making assumptions is a form of gambling.

1.6 Level 6 – Organ Changes



1.6.1 Detecting Organ Changes with Concentrations

The approach used to detect changes in average cells detects changes in organs as well.

Given this brief introduction to the big picture, we are ready to introduce the equations that we'll be using to operate within the theory structure being copied from biology.

CHAPTER 2

EQUATIONS

Solving biology involves translating published data into databases, databases into parallel complexities, and parallel complexities into equations that capture biology's rules and first principles. Equations and the routes thereto begin to define the boundaries of a complexity theory.

We will discover equations empirically by fitting simple and complex data types to linear regressions. When we collect data from biology, however, they become cloaked with uncertainty because they carry varying amounts of methodological biases and are subject to an uncertainty principle. Moreover, by taking biology apart, we change biology from a complexity to a simplicity and then view it through the lens of a theory structure convenient to the abbreviated data set. Understandably, the way we choose to look at our data determines what we will see.

If our goal is to solve biology, as it is here, then we want to see biology in its most pristine state. Since solving biology is largely an engineering problem, putting biology back together according to biology's rules and theory structure becomes a reasonable way to proceed. The easiest and most direct way of doing this is to detect, copy, test, and apply biology's rules as equations.

Chapter 2 identifies the principle equations that apply to each of six levels of complexity. Such equations apply to two broad categories of data, one the product of reductionism (biology taken apart) and the other of complexity (biology put back together).

2.1 Level 1 – Patterns

2.1.1 Data Pairs

The simplest unit of complexity consists of two named parts and one connection called a data pair - the first of several mathematical markers. We define it herein as a numerical ratio derived from the values of two parts (x, y), along with the names of the parts (a, b):

Data Pair Rule

$Data\ Pair = ax{:}\,by,$ wherein the

$$Ratio = \frac{x}{x}{:}\frac{y}{x} = 1{:}\frac{y}{x} \qquad (2.1)$$

Notice that by dividing both x and y by x, the first number in the ratio is set equal to 1. This tells us that one unit of x is associated with some number of units y.

2.1.2 Mathematical Markers

The alphanumeric strings of parts and connections create mathematical markers, wherein named parts (a, b, c, … , n) combine with numerical values (x, y, z, … , n) to form complex identifiers (ax:by:cz, … , n).

Mathematical Marker Rule

$$Ratio = \frac{x}{x}{:}\frac{y}{x}{:}\frac{z}{x}{:}\ldots\frac{n}{x} \qquad (2.2)$$

Notice that a triplet mathematical marker, which consists of three parts and two connections, can be arranged six ways: ax:by:cz, ax:cz:by , by:ax:cz, by:cz:ax, cz:ax:by, cz:by:ax. This turns out to be a particularly useful property in that we can expand the size of a given data set to include a much larger, all-inclusive set of searchable patterns.

If, for example, the rough-surfaced endoplasmic reticulum (RER) in a cell has a value of 2 m^2, the smooth-surfaced endoplasmic reticulum (SER) a value of 4 m^2, and the plasma membrane (PM) a value of 1 m^2, we get the following the string:

rer2:ser4:pm1.

By dividing all the values by x and removing the colons, we get the mathematical marker in standard form:

rer1ser2pm0.5.

By shifting from absolute values and concentrations to ratios, we can minimize both bias and variability. Although some baggage remains, there will be less of it (Bolender, 2016b)

Why is this new data type useful? By expressing published data as mathematical markers, even relatively small data sets can generate larger arrays of quantitative patterns. Moreover, collections of such patterns provide surprisingly effective approaches to diagnosis and prediction, while at the same time they allow us to see what biology can do quantitatively. By transforming an otherwise chaotic literature into universal biology databases consisting of highly organized and objective mathematical markers, we get to explore the literature as a unified whole. This allows us to treat the biological literature as one big experiment. Since everything depends on finding reproducible patterns, we're assembling a playing field in anticipation of the more challenging mathematical games we will be playing shortly.

Given such a resource, problem solving becomes an exercise in creating a database of mathematical markers either by applying filters to an existing database, or by combining markers from several databases. In all such cases, the strategy remains the same. By creating a complexity parallel to the one used by biology, we recruit biology to deliver the solutions we want. The better we are at copying biology, the better the result.

Mathematical markers will become increasingly important as we begin to simulate the complex interactions that occur within cells and organs. Although we consider here only the relationships of markers in one-dimensional strings, the universal databases containing such markers already suggest that biology is using n-dimensional networks of strings to run its business. In fact, this is exactly what we would expect to find in a highly adaptive and complex system, where everything interconnects.

2.1.3 Permutations

Recall the equation for calculating permutations (P), wherein n is the number of parts and r the sample taken [recall that 0! = 1]:

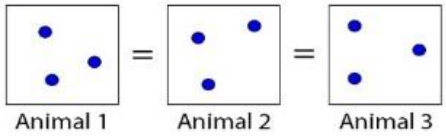
$$nPr = P(n,r) = \frac{n!}{(n-r)!} \qquad (2.3)$$

For example, 3 parts taken 3 at a time gives 6 permutations (3 x 2 x 1), whereas 10 parts taken 3 at a time gives 720 (10 x 9 x 8).

Why is this useful? The permutation equation allows us to reformat the biology literature – paper by paper – into a homogeneous data set consisting of universally compatible units of complexity (mathematical markers). This represents a critical step. Universal biology

databases solve two key problems. They condense the vast contributions of investigators into a common, searchable format and open lines of communication with biology by creating parallel complexities. Given their universal properties, mathematical markers are likely to become a favorite vehicle for big data and machine learning because they can provide optimized training sets.

2.2 Level 2 – Biochemical Homogeneity



A widely accepted principle of living systems is that organisms define themselves with relationships of structure to function. Perhaps the best-known application of this relationship was articulated by de Duve (1964, 1974), who gave us the postulates of biochemical homogeneity. They state that a marker enzyme is distributed uniformly at a single morphological location.

2.2.1 The Biochemical Homogeneity Test

Validating the postulates becomes a useful exercise for us here because success will allow us to dig deeper into multiple levels of biological complexity.

The design of the test is straightforward. First, we translate the postulate into an equation, evaluate the equation with published data, and finally demonstrate that the equation can predict expected outcomes using independent data sets.

The strategy behind the test may be the most interesting part. A successful outcome depends on our ability to devise a test so difficult that only biology can pass it. This means that the equations we propose must capture biological complexity. If not, we will not get the solution we want and expect. By bringing the notion of a parallel complexity into play, we align our interests closely with those of biology. The effect of such a strategy is that our job becomes that of an observer, leaving the theory structures and problem solving up to biology. The logic of such an approach seems obvious. Why would we want to construct a new wheel when a perfectly good one - with a massively impressive track record - already exists?

We want to test de Duve's homogeneity postulates literally and empirically. To this end, we must show that the activity of a membrane-bound marker enzyme - located at a unique cellular location – remains the same from one animal to the next when the marker enzyme activity is related to a square meter of the membrane surface area. Such a test will serve as a good example of how we can operate within the framework of a biological complexity, which, in this case, will consist of finding equations that capture biology's rules for producing membranes of the endoplasmic reticulum (ER).

2.2.2 The Structure-Function Rule (Part 1 of the Test)

Since de Duve used the rat liver to develop his homogeneity postulates, a similar source will be used for our test. Specifically, we will plot the surface areas of ER membranes - most of which come from hepatocytes (Blouin et al., 1977) - against the activities of marker enzymes bound to the ER membranes of these cells. The equation we need relates units of enzyme activity (U) to a unit of ER membrane surface

area (S). A general equation (Equation 2.4), which encapsulates biochemical homogeneity, defines the:

Structure-Function Rule as

$$y = f(x);\ f(x) = mx\ , \qquad (2.4)$$

where m is the slope of the line and x can be either a morphological or a biochemical value. Since the y intercept is equal to zero, the line passes through the origin (0,0).

The slope of the line (m) identifies a ratio either equal to U/S (an enzyme density, ED = U/S) or to S/U (a membrane density, MD = S/U), depending on the arrangement of the equation:

$$U = ED \times S \qquad (2.5)$$
$$U_j = \frac{U_j}{S_i} \times S_i\ ,$$

$$S = MD \times U \qquad (2.6)$$
$$S_i = \frac{S_i}{U_j} \times U_j$$

where i identifies a membrane, j a membrane-bound marker enzyme, U units of enzyme activity, and S membrane surface area.

If we reinterpret the data of an earlier study (Bolender, et al., 1978) by operating on the data one animal at a time (Bolender, 2017), we can evaluate equations 2.5 and 2.6 using an ER surface area and an ER marker enzyme (glucose-6-phosphatase); see Figure 2.1. This strategy generates the structure-function rules (Equations 2.7 and 2.8), as postulated by de Duve:

G-6-Pase-to-ER (Structure-Function Rule)

$$f(x) = 0.1677x \qquad (2.7)$$
$$y = 0.1677x\ ; wherein\ S = MD \times U$$

This equation calculates an ER surface area (y) from a membrane density (0.1677) multiplied by a glucose-6-phosphatase activity (x).

ER-to-G-6-Pase (Structure-Function Rule)

$$f(x) = 5.9625x \qquad (2.8)$$
$$y = 5.9625x\ ; wherein\ U = ED \times S$$

Alternatively, we can calculate glucose-6-phosphatase activity (y) from an enzyme density (5.9625) multiplied by an ER surface area (x). [Note: We can ignore the y intercepts displayed in the figures because they are very small (0.0011 and 0.0063).] As we'll see in Figure 2.4, similar results exist for other membrane-bound marker enzyme combinations.

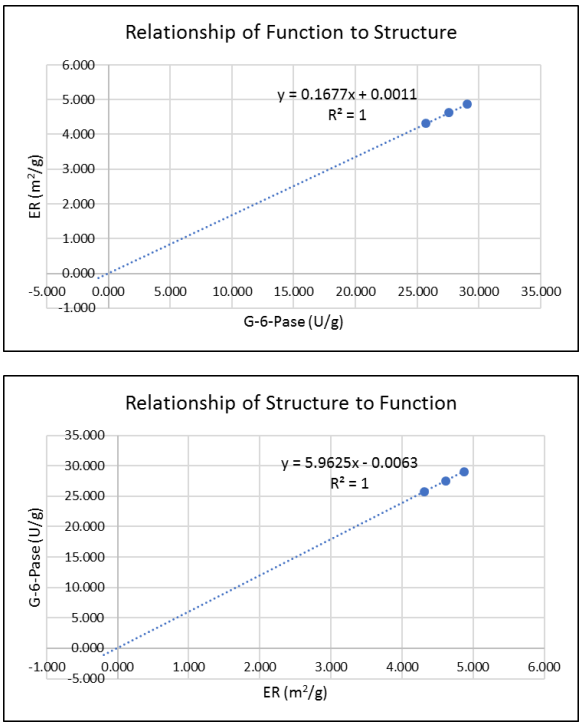


Figure 2.1 The plots identify a linear relationship between a structure (ER surface area) and a function (Units of a marker enzyme activity). The data of three animals fit the line with an $R^2 = 1$, which closely approximates the origin. Notice how we can use biological variation to our advantage. Animal to animal variation is allowing us to copy biology's rules with equations. Original data adapted from Bolender, et.al., 1978; From Bolender 2018.

The key point is that Figure 2.1 with its equations support the homogeneity postulates in that the data of three separate animals fall on the same line ($R^2 = 1$) and that the line essentially passes through the origin. In effect, equations 2.4 to 2.8 show how to capture the

rule biology uses to populate ER membranes of hepatocytes with molecules of glucose-6-phosphatase – in a control setting.

The enzyme density rule, which defines a ratio of function to structure, represents a concentration wherein the denominator is engineered in as a constant (set equal to 1 m^2), which prevents it from behaving as a variable. This gives us a common denominator within the cell that will allow us to map and integrate phenotypic and genotypic data.

2.2.3 Preparing for Part 2 of the Test

The second part of the biochemical homogeneity test is more challenging in that we need to show that the marker enzyme activity is uniformly distributed across the membranes of the ER – even though the ER contains two morphologically distinct ER subcompartments: rough- (RER) and smooth-surfaced (SER).

To this end, we'll need a few more equations along with some background information. Let's begin. We want to express an enzyme activity relative to an ER membrane surface area – instead of relating each measure to a gram of liver (Figure 2.1). This will allow us to look for differences in the amount of enzyme activity associated with 1 m^2 of ER.

The first equation we need is one that relates biochemistry (b) to morphology (m). Such a relationship defines a biochemical density (BD), which represents a concentration (b/m) and a mathematical relationship of structure to function (Equations 2.4 and 2.9).

Biochemical Density Rule

$$biochemical\ density\ (BD) = b/m\ . \quad (2.9)$$

Note that this rule adheres to the postulates of biochemical homogeneity (Equation 2.3) wherein:

Structure-Function Rule

$$f(x) = mx,$$

$$y = mx, m = \frac{y}{x}, and\ BD = \frac{U}{S} = b/m\ .$$

In turn, we can rewrite Equation 2.9 to accommodate our test by substituting marker enzyme activities (U) and ER membrane surfaces (S). This gives us the enzyme density rule (Equation 2.10):

Enzyme Density Rule

$$enzyme\ density\ (ED) = U/S. \quad (2.10)$$

This rule, which defines the enzyme density (ED) as a concentration (U/S), generates various outcomes wherein two knowns allow us to solve for one unknown:

$$ED = U/S \quad (2.11)$$
$$U = ED \times S$$

Recall that an enzyme density (U/S) *always* relates units of enzyme activity to 1 m^2 of membrane surface area. This gives us the constant reference we need to compare changes in EDs, to demonstrate reproducibility, and to predict biochemistry from morphology (Equation 2.11).

Similarly, we can express Equation 2.9 in morphological terms using the membrane density rule (Equations 2.12 and 2.13):

Membrane Density Rule

$$membrane\ density\ (MD) = m/b\ . \quad (2.12)$$

$$MD = S/U \quad (2.13)$$
$$S = MD \times U$$

Similarly, a membrane density (S/U) *always* relates membrane surface area to one unit of biochemical activity. This allows us to predict biochemistry from morphology. Since most published papers focus on either structure or

function, prediction becomes a powerful tool and often a game changer.

While equations 2.9 to 2.13 describe events occurring within the cell cytoplasm, enzyme densities can also extend their reach to organs by generating absolute values (equations 2.14 and 2.15). For example, enzyme densities (ED) predict total units (U_{total}) from total surface (S_{total}), whereas membrane densities predict total structure (S_{total}) from total units (U_{total}):

Absolute Value Rules

$$U_{total} = ED \times S_{total} \qquad (2.14)$$
$$U_{total} = U/S \times S_{total}$$

$$S_{total} = MD \times U_{total} \qquad (2.15)$$
$$S_{total} = S/U \times U_{total}.$$

Notice what we have done so far. By establishing quantitative relationships, we enable biochemistry and morphology to do things together that they cannot do alone. In other words, the strength of one becomes the strength of the other or the strength of one can compensate for the absence or weakness of the other. These are the types of relationships we want to build into the everyday operation of our theory structure.

Next, we need to look at the strategy behind enzyme densities. An enzyme density (ED) relates units of marker enzyme activity (U) to a unit (e.g., 1 m^2) of membrane surface area (S). The important piece of information is the way the concentration (U/S) is being defined. By dividing an enzyme activity by a membrane surface area given in m^2, the enzyme density will always be related to one square meter of surface area. This means that the biochemical reference (1 m^2) at any given point in time remains a rock-solid constant. The packing density of the enzymes in the membrane can change over time, but not the size of reference membrane. Why? When U is divided by S, S becomes equal to one square meter.

Why is this helpful? Using concentrations expressed in m^2 to detect a change ensures that the reference - the denominator of the ratio - remains constant. Biochemical concentrations (U/V), for example, referenced to a constant volume (V) can become grievously biased when the frequency of the cells filling the reference volume changes. Even a slight change in the average shape or volume of a cell type can result in the movement of surprisingly large numbers of cells into or out of this "standard unit" of reference volume. Standard in this context refers to a weight or volume, not to the contents. The contents – what's inside – behaves as a biological variable.

Changes in the contents of a gram or cm^3 of tissue occur routinely in experimental settings. But can such changes alter the outcome of an experiment? Yes, all too often. Indeed, this instability of the reference volume no doubt contributes generously to the ongoing crisis of precision, accuracy and reproducibility currently tormenting the biomedical sciences (Baker, 2016; Collins and Tabak, 2014; Engber, 2016; Freedman et al., 2015; Roth and Cox, 2015). By substituting concentrations based on EDs with a constant reference surface (1 m^2), we can avoid the mischief created by unstable reference volumes.

Now let's work through a calculation. If we relate the activity of an ER marker enzyme [glucose-6-phosphatase (G-6-Pase)] to the surface area of ER (both of which are related to a gram of liver), we can calculate an enzyme density for G-6-Pase:

$$ED_{G6Pase} = U_{G6Pase}/S_{ER} \qquad (2.16)$$

$$ED_{G6Pase} = \frac{29.031}{4.87} = 5.961 U_{G6Pase}\,/S_{ER}, \qquad (2.17)$$

where $S_{ER} = 1\ m^2$.

Equation 2.17 tells us that 1 m^2 of ER surface area carries 5.961 units of glucose-6-phosphatase activity.

2.2.4 Structure-Function Rule (Part 2 of the Test)

Now we can return to the problem we're trying to solve (the biochemical homogeneity test – Part 2). For glucose-6-phosphatase (G6Pase) and the ER, we have three pieces of information: the surface areas of the RER and SER and the activity of glucose-6-phosphatase – all related to a gram of liver. What we want are the individual enzyme densities for the RER and SER. This means that we need to write a balanced equation in two unknowns. Since we know from Equation 2.11 that units of enzyme activity (U) are equal to the enzyme density (ED) times the membrane surface area (S), we can write an equation with two unknown concentrations: ED_i, ED_j (Equation 2.16):

$$[S_i \times ED_i] + [S_j \times ED_j] = U_{total}\,, \qquad (2.18)$$

where

$S_i \times ED_i = U_i$ and

$S_j \times ED_j = U_j$, where

$m^2 \times \frac{U}{m^2} = U$, and

$U_i + U_j = U_{total}$.

For glucose-6-phosphatase and the ER compartments, we can rewrite Equation 2.18 to get the one we want for the test (Equation 2.19):

$$[S_{ser} \times ED_{ser}] + [S_{rer} \times ED_{rer}] = U_{G6Pase}\,. \quad (2.19)$$

Recall that:

$$U_{rer} + U_{ser} = U_{er}\,.$$

But how do we solve Equation 2.19 for the two unknown enzyme densities (EDs)? If we write two such equations populated with data coming from two different animals (Equations 2.20, 2.21) and solve them simultaneously, we can expect the resulting linear curves to intersect if both animals share the same enzyme densities for the RER and SER, but different surface areas and total enzyme activities.

$$[1.90 \times ED_{ser}] + [2.97 \times ED_{rer}] = 27.421 \quad (2.20)$$

$$[1.44 \times ED_{ser}] + [2.88 \times ED_{rer}] = 24.267 \quad (2.21)$$

Figure 2.2 supplies a graphical solution to equations 2.20 and 2.21, demonstrating that the lines intersect. Individual values for the ED_{ser} and ED_{rer} can be found by extending lines from the intersection point to the x and y axes. Note that the ED of the SER has a value of 5.77 and that of the RER 5.54.

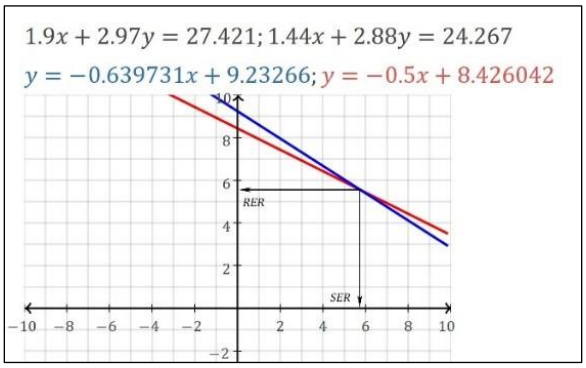


Figure 2.2 The ED of the SER is equal to 5.77, whereas that of the RER is 5.54 (Original data adapted from Bolender, et al., 1978).

Since the enzyme densities are almost the same (5.8 vs. 5.5), can we conclude that the ER membranes are biochemically homogeneous? To be on the safe side, let's run a quick t test.

If we write equations for three animals, pair them two by two to generate three estimates for the enzyme densities, then we have enough data to run a rough t test. The results shown in Figure 2.3 indicate that the 7% difference between the enzyme densities of the RER and SER membranes is significantly different (P = 0.0056).

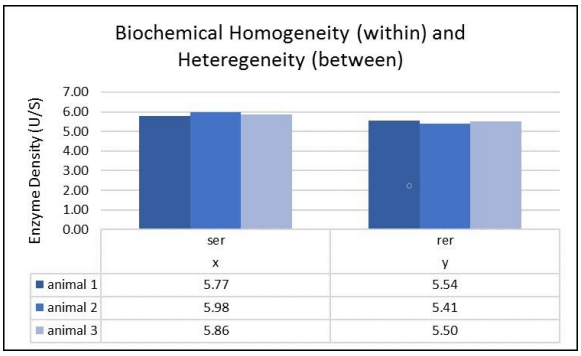


Figure 2.3 The enzyme density of the SER exceeds that of the RER by 7% (P=0.0056). A homogeneous ER appears to be the product of two underlying heterogeneities, unless we consider the ribosomes of the RER (Original data adapted from Bolender, et al., 1978; From Bolender, 2017).

But why would biology want to maintain a 7% difference between the RER and SER? It may not. Since ribosomes exist on the RER but not on the SER, their attachment sites are taking a small amount of the ER membrane surface area out of play – perhaps as much as 7% of the RER. This would seem to be the most likely explanation for the observed difference.

Why is this approach to problem solving with simultaneous equations of interest to us here? Detecting a biological difference of 7% with a P value of 0.0056 seems remarkable given the fact that the data came from multiple sources - animals, cells, membranes, and enzymes. Add to this the routine appearance of equations based on regression lines with R^2s equal to or close to one suggests we are dealing with a very exacting biology. Biology seems to have figured out that defining relationships of structure to function by rule improves its chances for success and survival.

2.2.5 Test Results

Both the data and the equations used for the test are consistent with the postulate of biochemical homogeneity as stated by de Duve (1964, 1974). It's important, however, to point out that the biochemical data reported originally as average data (Bolender et al., 1978) had to be adjusted to the membrane surface areas of individual animals (Bolender, 2017). We will explain in Chapter 4 how such results can be verified.

2.2.6 Insights from Level 2 Complexity

The equations of Level 2 demonstrate that problems related to complex relationships of structure to function need equally complex solutions. This means that solving biology involves identifying a new generation of equations capable of capturing and vetting the many rules that define biology.

Figure 2.4, for example, includes a short list of rules (equations) defining such relationships for several membrane organelles in control hepatocytes. The table allows predictions both ways – structure to function and function to structure. Moreover, duplicating such equations across publications becomes a test of reproducibility.

Rules for Membrane Organelles (Hepatocyte)

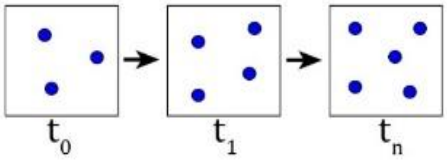
INPUT X		PREDICT Y			
X	UNITS	Y	UNITS	EQUATION	R^2 =
S(ER)	M^2/G	G-6-Pase	U/G	Y = 5.9625X	1
G-6-Pase	U/G	S(ER)	M^2/G	Y = 0.1677X	1
S(ER)	M^2/G	Esterase	U/G	Y = 46.657X	1
Esterase	U/G	S(ER)	M^2/G	Y = 0.0214X	1
S(ER)	M^2/G	NADPH-CCR	U/G	Y = 0.9829X	1
NADPH-CCR	U/G	S(ER)	M^2/G	Y = 1.0174X	1
S(OMIM)	M^2/G	MAO	U/G	Y = 0.6310X	1
MAO	U/G	S(OMIM)	M^2/G	Y = 1.5848X	1
S(IMIM)	M^2/G	CYOX	U/G	Y = 6.8150X	1
CYOX	U/G	S(IMIM)	M^2/G	Y = 0.1467X	1
S(PM)	M^2/G	5'NUC	U/G	Y = 26.267X	1
5'NUC	U/G	S(PM)	M^2/G	Y = 0.0381X	1

Figure 2.4 Using enzyme and membrane densities to represent biochemical homogeneity, we can predict enzymes from surfaces and surfaces from enzymes. Original data adapted from Bolender, et al., 1978; From Bolender 2017. [Note: the biochemical data come from

tissue homogenates (E + N), collected as E (extract) and N (nuclear) fractions.]

Figure 2.4 raises a fundamental question triggered by a central theme running through the literature of molecular biology. Are these equations restricted to rat livers, or do they have a broader distribution? Since it appears that livers of similar animals, such as rats, mice, and humans, share many of the same genes, might they also be sharing equations akin to the ones shown in Figure 2.4? Will, for example, the similarities and differences in the networks of equations needed to capture and simulate liver phenotypes tell us more about the principles underlying gene expression than we might otherwise learn by focusing our attention on just genetic constructs? Is it even possible to uncover the deeper secrets of genomes without giving more attention to their phenotypes?

2.3 Level 3 – Organelle Changes



The pattern emerging from the narrative thus far is that relationships of structure to function are helping us to understand how the biology is playing its complexity game.

Level 3 complexity ups the ante by creating playing fields of greater complexity wherein biochemical homogeneities (expressed as enzyme densities) change over time. This identifies the next problem to work on.

Let's review what we know so far. Plots of structure to function produce $R^2 = 1$ equations (Figure 2.4) when multiple animals share the same enzyme density (e.g., Figure 2.1). This suggests that we start to work on the change problem by calculating enzyme densities at each of several experimental timepoints to see if they change in an orderly way.

2.3.1 Structure-Function Change Rule

Figure 2.5 plots enzyme densities for three ER marker enzymes from hepatocytes exposed to a drug (phenobarbital) for five days. Notice that the enzyme densities (U/S) changed, but at rates unique to each enzyme. In fact, the hepatocytes control the rates at which new enzymes become incorporated into their ER membranes with near laser-like precision – the regression lines carry R^2s = 1 or ≈ 1.

The equations shown in Figure 2.5 identify a general rule for biological changes occurring in cells:

Structure-Function Change Rule

$f(x) = mx + b$, where (2.22)

$y = f(x), m = ED, and\ x = time.$

For phenobarbital (100 mg/kgbw/d/5 days), the following rules quantify the biological changes occurring in the ER membranes of hepatocytes.

***Structure-Function Change Rule**: Cytochrome P450 + ER Surface Area*

$$ED_{U_{cytoP450}/S_{er}} = 0.4223x + 0.9187 \quad (2.23)$$

***Structure-Function Change Rule:** n-demethylase + ER Surface Area*

$$ED_{U_{n-demethylase}/S_{er}} = 0.1470x + 0.4017 \quad (2.24)$$

***Structure-Function Change Rule:** NADPH-cytochrome c reductase + ER Surface Area*

$$ED_{U_{nadph-ccr}/S_{er}} = 0.0018x + 0.0069 \quad (2.25)$$

Notice that biological complexity involves the embedding of rules (equations) in rules (equations). The two sets of rules are

expressed simultaneously: one defines the local EDs (Equation 2.4) whereas another determines the slope of curve on which the EDs sit (Equation 2.22). A given point defined by one rule (Equation 2.4) generates a second rule (Equation 2.22) when a change occurs.

This tells us that by not knowing the rules that govern a change, we are reduced to running an experiment in a "black box" wherein a cast of many variables perform together to deliver a highly opaque result – one concentration or one absolute value. We get a result but don't have a clue as to how biology solved the problem. Ignoring biology produces another shortcoming. We wouldn't know, for example, that our results are several levels of complexity removed from where we need to be to make mathematical connections within the phenotype or to work our way back to events occurring in the genome. In effect, we become blindsided by our outdated experimental model.

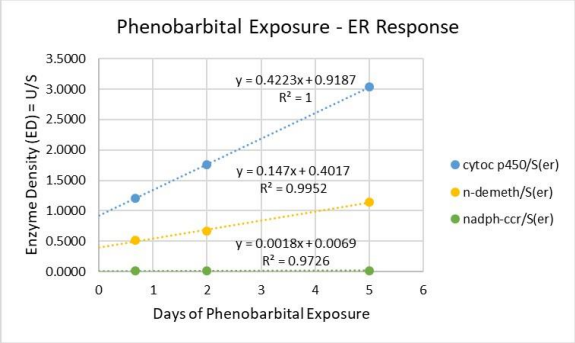


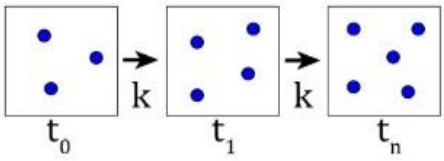
Figure 2.5 The enzyme densities of ER membranes in hepatocytes display a highly orchestrated response to the drug phenobarbital. (Original data adapted from Stäubli et al., 1969; From Bolender, 2018).

The equations in Figure 2.5 identify linear growth curves that define changes in the cytoplasmic membranes (ER) of hepatocytes. What triggers such a complex but highly orchestrated series of events? How, for example, does biology assemble and execute the recipes needed for such changes? Where are these change algorithms located? How do we find them?

Figure 2.5 shows us that the enzyme densities continue to change over time in response to the drug – each one at a different rate. This means that duplicating an experimental time point represents a far more problematic undertaking than duplicating an equation. The same argument applies to looking for evidence of reproducibility in the literature.

Thus far, all we know is what's happening locally at the level of the ER membranes. Figure 2.5, however, contains a pivotal clue that takes us to the next level of complexity. The R^2s tell us that we can use enzyme densities – because they are concentrations - to estimate rate constants for ER membranes of hepatocytes as they exist *in situ*. But how do these new rate constants compare to those estimated the usual way *in vitro*? Once again, we're positioning ourselves to ask questions fundamental to understanding the complexity of membrane changes occurring in living cells.

2.4 Level 4 – Rates of Change



Rate equations define the speed at which a concentration changes over time. When plotted against time, the shape of the resulting curve identifies the order of the reaction. The linear plot shown in Figure 2.5, for example, indicates that we are dealing with a change analogous to a zeroth order reaction.

2.4.1 Rate Equation

The integrated rate law for a zeroth order reaction is given as:

$$[A] = [A]_0 \pm kt\,, \qquad (2.26)$$

where [A] is the concentration a given time, $[A]_0$ the concentration at time 0, k the rate constant, and t the given time. Since Equation 2.26 describes a straight line (y = mx + b), $[A]_0$ becomes the y intercept and k the slope (m). Recall that a minus sign (-) identifies a negative slope, whereas a plus sign (+) a positive slope.

For example, to calculate the missing value for day 3 in Figure 2.5, we can substitute the concentrations of Equation 2.22 with the enzyme densities (ED) for Cytochrome P 450 (Note that the original day 3 becomes day 2 here because day 1 is equal to day 0 in the rate equation):

$$[ED]_3 = [ED]_0 + kt_3 \qquad (2.27)$$

$$[ED]_3 = [1.341]_0 + 0.4223 * 2$$

$$[ED]_3 = 2.1856.$$

Note that the slopes of the lines in Figure 2.5 tell us the rate (k) at which enzyme molecules are being added to the ER membranes of hepatocytes exposed to phenobarbital. The more familiar application of enzyme kinetics would include a biochemical test done *in vitro* wherein the enzyme data are related to a mg of protein or to a gram of liver.

2.4.2 Consequences of Disconnecting Structure from Function

If instead of plotting changes in enzyme densities, what would happen to the rate constants if we plot the changes in enzyme activities and membrane surface areas separately?

Now the results tell us an entirely different story. Figure 2.6 shows that the rate constant for cytochrome P450 compared to its change in enzyme density (Figure 2.5) differs by more than 10-fold (0.4223 vs. 4.5117).

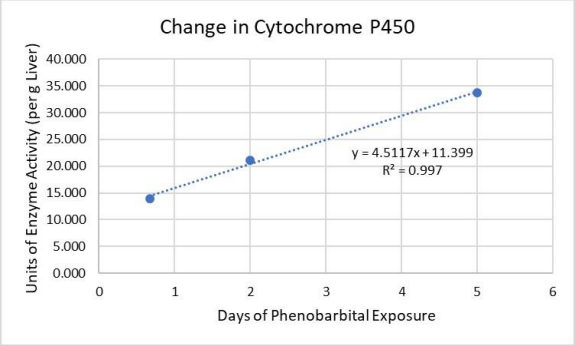


Figure 2.6 The rate constant (k) for cytochrome P450 becomes 4.5117 when calculated with biochemical data alone (Original data adapted from Stäubli et al., 1969).

Plotting the changes in membrane surface areas gives a similarly disappointing result (Figure 2.7).

For a zeroth order event, the rate constants for the changes in ER membrane surface area display a negative slope (Figure 2.7) with data points widely scattered about the regression line (R^2 = 0.5299).

Figures 2.6 and 2.7 illustrate the typical risks involved when running black box experiments. In addition to getting inconsistent results, we would be hard pressed to explain why.

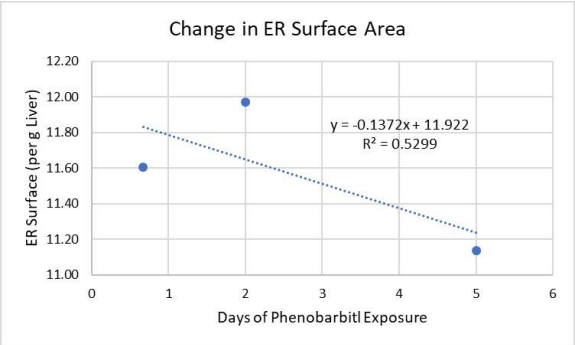


Figure 2.7 When ER surface area is plotted against time, the rate constant for the ER membranes displays a negative slope with an R^2 = 0.5299. Since the cells are filling up with ER membranes (mostly SER), they get bigger

and fewer of them can fit inside a gram of liver. In fact, after five days of exposure to phenobarbital, hepatocytes have about twice as much SER surface area compared to the control (Original data adapted from Stäubli et al., 1969).

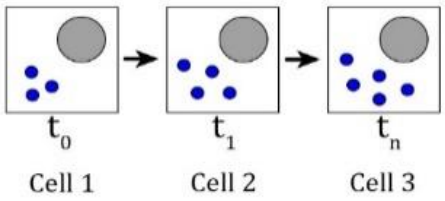
What are Figures 2.5, 2.6, and 2.7 trying to tell us? When we separate structure from function, we jettison the safety net provided by biology's rules and end up with inconsistent results we can't explain.

Although Figures 2.5, 2.6, and 2.7 identify the hidden cost of disconnecting structure from function, what additional price might we be paying for removing the membranes, enzymes, and enzyme densities from their hepatocytes? How do we reconnect the many parts we have marginalized by removing them from their original setting in a cell?

In contrast, the complexity captured with equations at levels 2, 3, and 4 provide detailed information attached to a quantitative foundation. The phenotype with its ability to express a change as a relationship of structure to function can also offer the structural support needed to interpret genomic data, which derive largely from biochemical data (see, for example, Alberts et al., 2014).

Next, we will need to figure out how to put the complexities of levels 2, 3, and 4 back into the cells where they exist.

2.5 Level 5 – Cell Changes



2.5.1 Detecting Cell Changes

There are two ways of detecting how a part changes in a cell. They include calculating average cell data or always relating the data of organelles or other cell compartments to the same number of cells. Usually, we can estimate average cell data by dividing a total organ value by the total number of cells contained therein. Such a route to average cell data, unfortunately, does not work for the liver. Why? The stereological methods used to estimate cell numbers rely on counting nuclei – not cells. Liver hepatocytes have several categories of nuclei: diploid (2N), polyploid (4N+), and binucleated. Moreover, the frequency of binucleated cells in the liver changes routinely in response to metabolic demands (hepatocytic nuclei can undergo both fission and fusion). This means that choosing the average cell option for the liver is a non-starter because it only adds uncertainty to the results.

2.5.2 Detecting Cell Changes with Concentrations

The remaining option for the liver hepatocytes is to collect data from the same number of cells. Let's begin by reviewing what we know so far.

Hepatocytes respond to the drug phenobarbital by synthesizing new ER membranes capable of metabolizing the drug. Consequently, the hepatocytes must increase their volumes. Such volume increases, however, will quickly destabilize the effectiveness of a reference based on a gram or cubic centimeter of liver, both of which are routinely used as references for morphological and biochemical data. Why? The number of cells that can be packed into a cm^3 of liver depends of the size of the cell - the bigger the cells, the fewer are needed to fill the reference volume. [Note: a cm^3 of rat liver weighs about 1.065 grams; $\rho(density) = weight/volume$.]

This means that experimental changes indexed to a gram of liver will be the result of changes to the cytoplasmic compartments that influence cell volume and/or to changes in the number of hepatocytes needed to fill a gram (or cm^3) of liver. This ambiguity (multiple variables in play) sets a trap for the unsuspecting investigator.

What is this telling us? By following changes in just biochemical or organelle concentrations, we ignore the larger complexity of a biological change and run the risk of getting incomplete and misleading results. This also tells us that concentration data contribute generously to the chaotic state of the biology literature.

One way of resolving this problem of ambiguity – being brought about by multiple, simultaneous changes - is to maintain the same number of hepatocytes filling a cm^3 of liver – even when the cells increase or decrease in volume. To do this we will allow the original control cm^3 of liver hepatocytes to swell (hepatocytes get larger) or shrink (hepatocytes get smaller) – as the need arises. This temporary distortion of reality – a flexible cm^3 – relies on Equation 2.28 to recover reality and to mitigate our concentration problem.

2.5.3 Concentration Correcting Equation Corrected (CCC)

The equation we want is one that keeps the number of hepatocytes associated with a gram (or cm^3) of liver the same (as it originally existed in the controls) – even when an experimental setting causes the hepatocytes to become bigger or smaller. To this end, we need to assume that the number of hepatocytes in the liver remains constant and that all the changes in liver weight (or volume) are attributable to the hepatocytes. To introduce a correction based on just changes in hepatocytes, however, all contributions coming from the rest of the liver – the extrahepatocytic compartments - must be considered. This results in a concentration equation corrected twice:

$$CCC_{(t_i)} = C(t_i)\, x \frac{W_L(t_i) - W[EHS(t_0)]}{W_L(t_0) - W[EHS(t_0)]}\ , \qquad (2.28)$$

where $CCC_{(t_i)}$ is the corrected concentration at time i, $C(t_i)$ the uncorrected concentration at time i, $W_L(t_i)$ the weight of the liver at times i and 0, and $W[EHS(t_0)]$ the weight of the extrahepatocytic space (EHS) at time 0 (t_0). The equation assumes that the EHS remains constant throughout the experiment. Note that the EHS accounts for about 6.8% of the total liver volume (Weibel et al., 1969).

How effective was Equation 2.28 in correcting the concentration data? Figure 2.8 compares the concentrations before and after the corrections of Equation 2.28 were applied to an ER marker enzyme - cytochrome P450. After 5 days of exposure to phenobarbital, the enzyme activity (expressed as a concentration) showed a 3.6-fold increase when related to a gram of liver, but a 5.6-fold increase when related to a gram of liver containing a constant number of hepatocytes – a 54% difference. Recall that bigger differences usually translate into better P values, which determine the strength of significant differences.

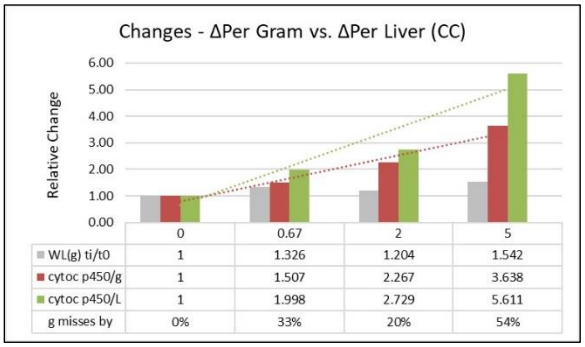
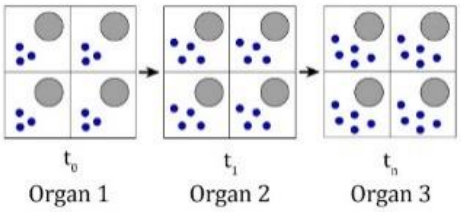


Figure 2.8 Relating biochemical data to a gram of liver, as it actively loses large numbers of hepatocytes, can underestimate changes in enzyme activities by a substantial amount (Original data adapted from Stäubli et al., 1969; From Bolender, 2018).

2.5.4 Insights from Level 5

The advantage of equation 2.28 is that it can remove the ambiguity associated with both morphological and biochemical concentrations in experimental settings – without the need to count cells. Although designed originally for liver hepatocytes, it should apply as well to other cells and organs. In such cases, the results can be checked by counting cells with the fractionator (Gundersen et al., 1988) and estimating the average cell changes directly.

2.6 Level 6 – Organ Changes



The least ambiguous way of interpreting a morphological change is to relate the data to an organ or to an average cell. At complexity level 6, we now have two new options for estimating a total biological change in hepatocytes without having to resort to hierarchy equations or cell counts.

2.6.1 Detecting Total ER Surface in the Liver (Method 1)

The advantage of Equation 2.28 is that it corrects the concentration data of both morphology and biochemistry by equating changes in average hepatocytic volumes or weights (Figure 2.24) to changes in liver volume or weight (Figure 2.25):

$$CCC_{hep} \Rightarrow Change\ per\ Av.Hepatocyte \quad (2.29)$$

$$CCC_{liver} \Rightarrow Change\ per\ Liver \quad (2.30)$$

$$\Delta CCC_{avhep} = \Delta CCC_{liver} \quad (2.31)$$

The disadvantage, of course, is that equation 2.28 may not work as well when applied to other organs with more heterogeneous cell populations or when the numbers of multiple cell types undergo substantial changes.

2.6.2 Detecting Total ER Surface in the Liver (Method 2)

Recall that we can also get information about changes in average cells by estimating the total amount of ER membrane in the liver with equation 2.32:

$$f(x) = S_i/U_j \times U_{total} \quad (2.32)$$

$$S_{total} = MD \times U_{total}\ .$$

Where i is the name of a membrane compartment and j that of a membrane-bound marker enzyme.

If, for example, we start with the activity of cytochrome P450 and the membrane density (MD), we can use Equation 2.15 and the available data to estimate the total ER surface area in the liver. After five days of exposure to phenobarbital, the liver contains 68.4 m^2, almost double the amount found in the control:

$$S_{total} = MD \times U_{total} \quad (2.33)$$

$$S_{er,liver} = MD \times U_{cytoP450,liver}$$

$$S_{er,liver} = (0.3298\ m^2/U) \times 207.3336\ \mathrm{U} = 68.3840\ \mathrm{m^2}$$

Notice that the primary difference between the two methods is that they operate under two different theory structures. In method 1, morphology and biochemistry work alone, but require two assumptions. In method 2, morphology and biochemistry work together to generate the same results without having to rely on the assumptions of method 1. The lesson? Copying biology tends to be an easier and safer approach to problem solving.

Be careful. Both methods 1 and 2 assume that biochemical data collected from microsomes faithfully represent total liver values.

Now that we have worked through the fundamentals of approaching biology as a complexity, we can look at some of the experimental applications. Bear in mind that we will be updating the ground rules for running literature experiments to accommodate biological complexity.

CHAPTER 3

VISUALIZING COMPLEXITY

Solving biology requires a deliberate process, one that involves working out relationships of parts to connections to patterns to equations - all bridging multiple levels of complexity. Seen from afar, the magnitude of such a task may at first appear daunting, but up close one quickly discovers that this is the way biology actually works.

Let's put into perspective what we are trying to do. Complexity in physics and chemistry derive from the elements of the periodic table, whereas biology must cope with both the periodic table and one of its own making – a vast table of genes. If biology has already figured out how to deal with all this complexity, then the first thing we must do is figure out what biology is doing. Then we can discover how.

In this chapter, we'll see how solving biology represents an iterative process of finding clues that lead to more clues that eventually lead to solutions. The persistent challenge, of course, includes learning how to recognize the clues when they appear. It's an acquired skill.

The process of moving from one level of complexity to the next often involves making the transition from one database to another. The first level of complexity requires large data sets because the goal is to gain a mathematical footing within a literature database populated with just parts data. Almost immediately, reproducible patterns point to an underlying order. Such patterns supply the clues that become the stepping stones leading us from one level of complexity to the next.

3.1 LEVEL 1 – PATTERNS

Although the equations of Chapter 2 showed that biology relies on relationships of structure to function to define its complexity, much can be learned from the way biology orders its parts with ratios. Level 1 patterns, for example, show considerable promise for building large scale diagnostic and predictive systems for healthcare, for identifying phenotypic responses to genetic events, and for upgrading the reliability of experimental methods.

Since we already know from molecular biology that animal species display remarkably similar genomes, it follows that we can expect to find similar phenotypic patterns within and across species. Keep in mind, however, that such patterns can change throughout life as they adapt to a wide range of internal and external influences. Although we still don't know enough to interpret many of the changes that appear, the patterns nonetheless generate more than enough clues to suggest where we might want to go next.

3.1.1 RATIOS

Recall that the simplest unit of biological complexity consists of two named parts (a, b) and one connection (x:y) – the *data pair* (ax:by).

We define it herein as a numerical ratio derived from the values of two parts.

Expressing published data as ratios solves two pivotal problems. It creates a uniform and universal sets of searchable patterns, while at the same time it removes reference system instabilities that weaken morphological and biochemical data. For example, dividing one concentration (y) by another (x) eliminates the offending reference volumes (v_{ref}) by simply cancelling them out. The result is a dimensionless ratio:

$$\text{Concentration y}/\text{Concentration x} =$$

$$\frac{(4\text{y}/\,v_{ref\,i})}{(2\text{x}/v_{ref\,i})} = \frac{4\text{y}}{2\text{x}} = \frac{2\text{y}}{1\text{x}} = 1\text{x}{:}\,2\text{y}, \qquad (3.1)$$

where $v_{ref\,i} = v_{ref\,i}$.

The biology literature stores vast amounts of concentration data. Stereological estimates, for example, generate volume (v/v), surface (s/v), length (l/v), and numerical (n/v) densities – all of which exist as concentrations with potentially unstable reference volumes. Similarly, biochemical data expressed as concentrations (an activity or amount of a constituent per mg protein or g of tissue) can display similar instabilities (e.g., Figure 6.6).

By plotting one set of parts against those of another, it quickly becomes apparent that ratios serve as the backbone of biological complexity. Such sets, which include data pair libraries, compare controls to controls and controls to experimentals. By retaining data pairs that produce regression lines with R^2 close to one (e.g., 0.999), we discover that similar and different parts coming from a wide range of species can share the same regression equation. Figure 3.1, for example, shows that the relationship of the endoplasmic reticulum to other cytoplasmic organelles can display a high degree of order, as indicated by an $R^2 = 0.99$.

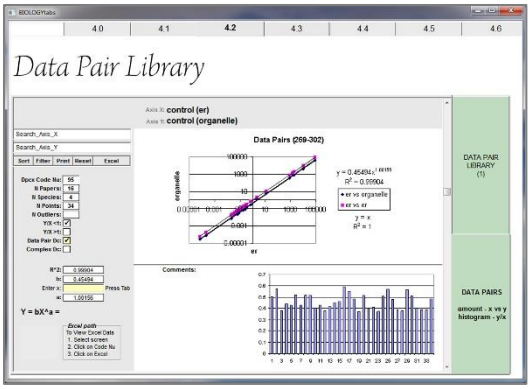


Figure 3.1 A high degree of order can be found in the relationship of the endoplasmic reticulum to other cell organelles (From Bolender, 2004).

Moreover, selecting data pairs with regression lines becomes an effective strategy for extracting patterns from clumps of data points. Figure 3.2, for example, detects unsuspected relationships between mitochondrial and Golgi membranes.

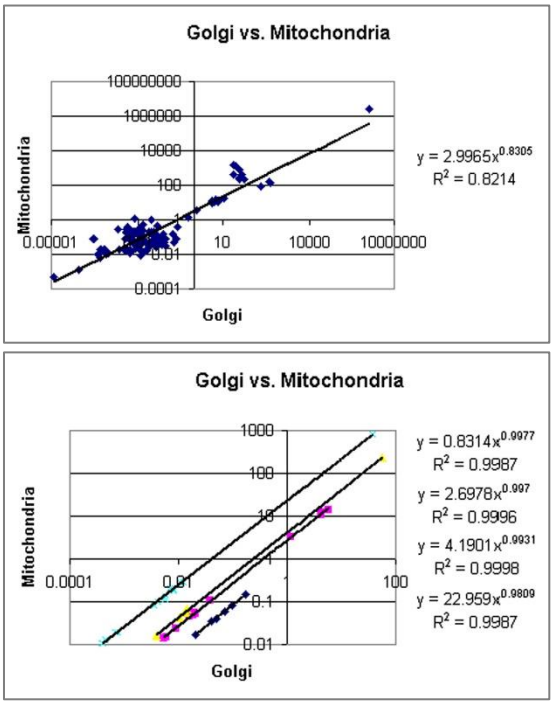


Figure 3.2 A repertoire plot shows the relationship of mitochondria to Golgi. Top: Typically, such comparisons display data clumps with weak correlations. Bottom: When fitted to repertoire equations by removing outliers (defined here as points not on or close to the line), the data clumps unfold into a set of parallel lines with R^2s close to 1.0 (From Bolender, 2004).

One of these figures, however, tells a particularly interesting story. When we plot mitochondria against other cell organelles and fit them to a regression line, the data pairs selected thereby produce a histogram with discernable steps (Figure 3.3).

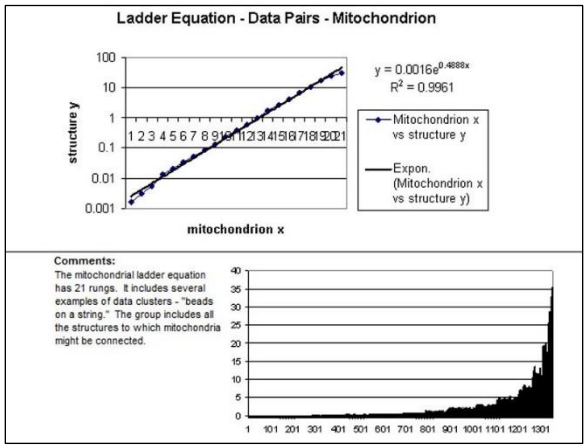


Figure 3.3 When a ladder equation plots the mitochondrion against other organelles and is fitted to a regression with an $R^2 = 0.99$, the corresponding histogram of the data pairs display distinct steps (From Bolender, 2004).

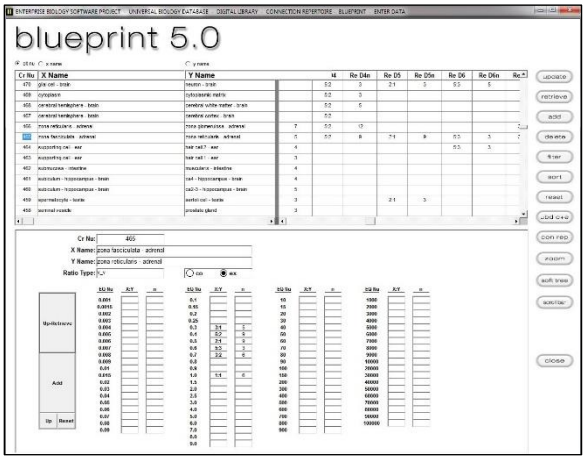
3.1.2 Decimal Ratios

The histogram of Figure 3.3 suggests that the relationship of one part to another describes a digital (noncontinuous) rather than an analogue (continuous) distribution. In effect, Figure 3.3 provides an important clue. If we use the permutation equation (2.2) to expand the parts data stored in the stereology literature database and then assign the resulting ratios to discrete decimal bins (ranging from 0.0001 to 100,000; see Figure 7.13), published data become optimized for finding quantitative patterns.

By expanding (taking permutations) and standardizing published data (forming data strings), we end up with a working model for a *Universal Biology Database,* one designed specifically to ferret out quantitative patterns in the biology literature. This switch from analogue to digital triggers a new strategy, one that consists of assembling new databases from the biology literature designed to address specific problems. Such a strategy continues to be highly effective in identifying and resolving a wide range of complex issues.

The first task of the new universal database was to convert an undisciplined collection of decimal ratios into a biological blueprint, one that could provide clues to the way biology organizes its parts.

Figure 3.4 shows the data entry screens used to populate the blueprint database and Figure 3.5 illustrates a typical search using a query by example (QBE) interface. In short, the blueprint database plays a key exploratory role in that it provides empirical evidence for the existence of inter- and infraspecific similarities across a wide range biological parts and phenotypes. Such a finding dovetails with numerous reports in the literature of molecular biology that describe extensive sharing of genes across species (Alberts et al., 2014; Lodish et al., 2016).



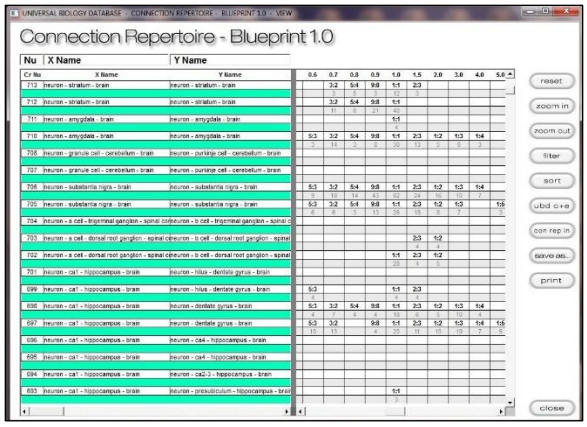


Figure 3.4 Top: Data entry consists of recording all the connections (as ratios) associated with a given pair of parts. Bottom: The blueprint database documents the distribution of data pairs, ratios, valences, and frequencies (From Bolender, 2006).

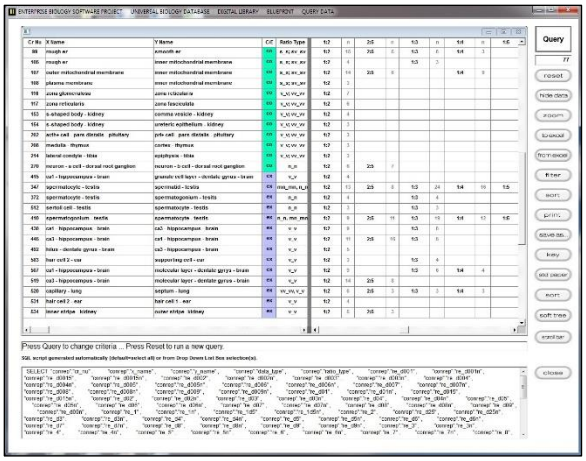
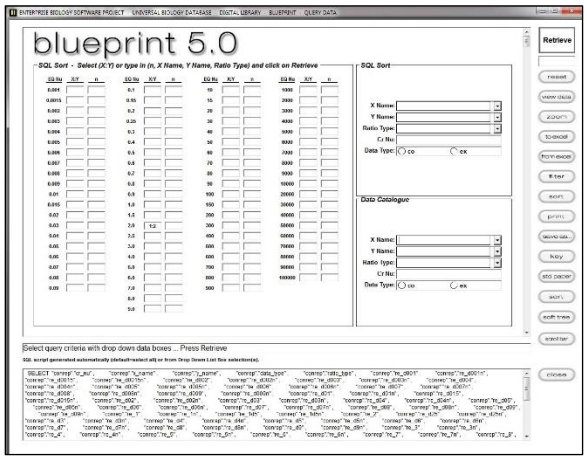
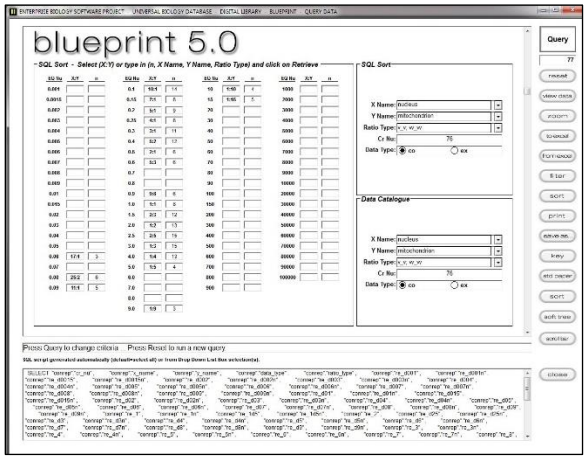


Figure 3.5 Top: The SQL query selects the x:y ratio of 1:2. Middle: Clicking on the Query Button sends the request to the database, which returns the information requested; 77 examples of 1:2 rations were found. Bottom: Individual data pairs are listed along with the extent of their reproducibility (the number of duplicates) in the database. The SQL script shown at the bottom of the screen includes directions the database uses to find the 77 examples of 1:2 ratios (From Bolender, 2006).





3.1.3 Universal Biology Database

A key advantage of having a stereology literature database is that we can use it to generate new databases specific to a given problem – quickly and easily. A universal biology database translates the original parts data of the literature into decimal ratios and mathematical markers. These dimensionless ratios help to resolve the long-standing reference system problems (see concentration trap; Figures 7.2, 7.3, and 7.4) by converting both concentrations and absolute values into a standardized set of alphanumeric strings. This proves to be helpful. Instead of having to deal with data scattered across thousands of papers, research data are now sitting side by side in the same place and in the same format. Given such an arrangement, they can work together.

When expressed as decimal ratios, this new data type also fits comfortably with the view that complex adaptive systems evolve toward

the edge of chaos, a transitional zone between order and chaos where complexity is maximal. Data collected from the biology literature display a high degree of exactness (order) inconsistent with the load of methodological bias they habitually carry. By assigning such data to decimal bins, the exactness of the published data becomes slightly less exact, thereby moving them toward or onto the edge of chaos. At this edge, patterns flourish, and emergent properties appear (Walthrop, 1992, Kauffman, 1995).

An obvious application of the universal database is to explore the consequences of genetically modifying organisms. With new directions in molecular biology being forged with CRIPSR and imRNA, where are we headed? Will the root changes we are introducing into biology's complexity become analogous to the triggering butterfly effect of chaos theory?

A hint as to the implication of genetically engineering organisms, for example, comes from a study of cell ratios in the lateral geniculate nucleus of 58 isogenic strains of mice (Seecharan et al., 2003; Bolender, 2005). Plotting data pairs derived from the original data set (cell counts) produced a cloud of poorly correlated data points ($R^2 = 0.03$). However, the data pairs of the cloud unfold into 14 equations displaying R^2s greater than 0.9. These results – shown in Figure 7.7 - tell us that 58 isogenic strains of mice have at least 14 unique ways to build a lateral geniculate nucleus – starting with three cell types. Such unexpected results suggest that the butterfly's wings have already started to beat.

3.1.4 Data Triplets

At this point, we're ready to scale up the information content of the original stereology literature database by increasing the number of variables in the data ratio from two (ax:by) to three (ax:by:cz).

Consider three named parts, a, b, and c with values x, y, and z. Three parts taken two at a time gives six data pairs: a:b, a:c, b:a, b:c, c:a, and c:b. When combined with their ratios, we get six mathematical markers - ax:by:cz, ax:cz:by, by:ax:cz, by:cz:ax, cz:ax:by, and cz:by:ax – identified herein as triplets.

Figure 3.6 shows how generating all possible data ratios has a strong multiplier effect. Five original points taken two at a time give 20 data pairs, 10 points give 90, and 25 points give 600. Taking n parts 3 at a time to form data triplets takes it up a notch: 5 points now give 60 triplets, 10 give 720, and 25 give 13,800.

Quadruplets, however, can quickly push the numbers well beyond the ability of the current technology to manage them. An Excel (Microsoft, Redmond, WA) spreadsheet, for example, accepts only about 1.2 million rows of data.

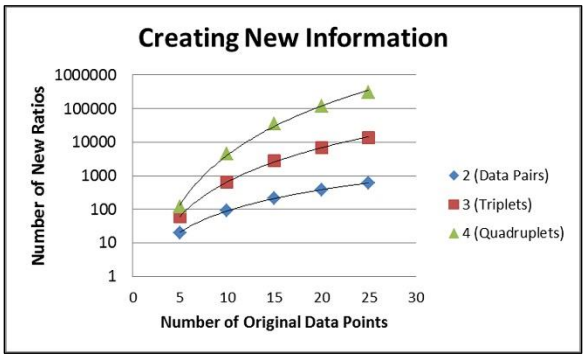


Figure 3.6 A data set can be expanded by taking all possible permutations taken two at a time (Data Pairs), three at a time (Triplets), and four at a time (Quadruplets). The plot illustrates what happens to the data coming from a single paper. Although roughly 46,000 data pairs and 850,000 data triplets were used to populate the universal databases, the number of quadruplet markers exceeded 15,000,000 (From Bolender, 2016a).

3.1.5 Organism Codes

Although we can translate our published data into patterns with triplet ratios, how do these

patterns fit into biology's scheme? In fact, the fit seems to be remarkably good. For example, we can use organism codes to represent the parts of a cell either as a 1D string or a 2D network of ratios. In Figure 3.7, the organelles of the gastric parietal cell form a 2D network of interconnections. This arrangement of parts – expressed as ratios – identifies the rules biology is using to define a normal, resting parietal cell. By translating biology's rules of design into visual patterns, we see how complexity is being expressed.

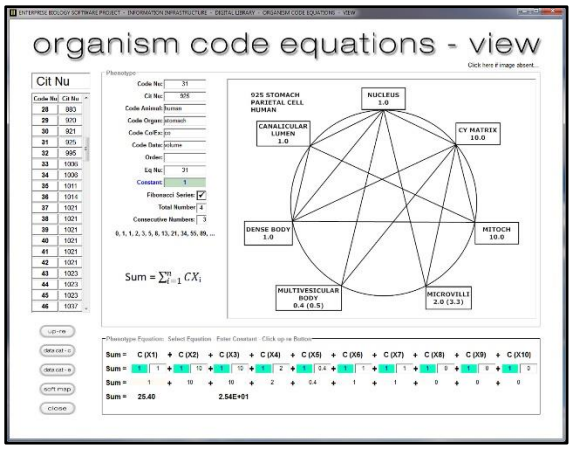


Figure 3.7 Organelles of the parietal cell in the human stomach display multiple connections. Moreover, the connections can be combined to form a string of ratios (lower panel), which reflects the biological rule (code) for constructing parietal cells - nuc(1) : cyma(10) : mito(10) : mivi(2) : mvb(0.4) : db(1) : calu(1) (Original data adapted from Aase et al., 1976; From Bolender, 2010).

Do biology's rules change? Yes. Can we influence such changes? Yes, again. Figure 3.8 illustrates the patterns displayed by the human hippocampus in health (top), alcoholism (middle), and Alzheimer's disease (bottom). Notice that both the ratios and the connections can change. In Alzheimer's disease, multiple parts of the hippocampus have lost their ability to network with all the other parts except for the presubiculum. In effect, the brain shuts down communication by dropping connections based on ratios. The normal recipe disappears.

What clue jumps out from Figure 3.8? Alzheimer's disease changes the parts and connections of the brain quantitively. The clue suggests that such changes might be diagnostic for this disorder, or even for disorders in general. If we take the bait and follow the clue, what happens? We end up with a new diagnostic test for disorders of the brain (Figure 3.23).

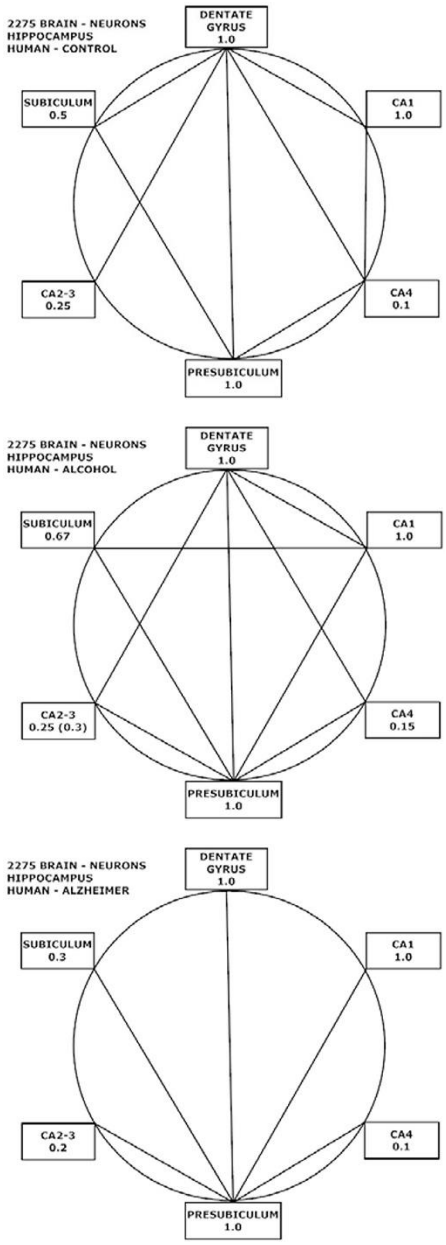


Figure 3.8 Organism codes - based on triplets - characterize the hippocampus in health (control: top) and disease (alcohol: middle, Alzheimer: bottom). Notice how the

triplets can detect changes in the complexity of the hippocampus, using the relationship of parts to connections (Original data adapted from Harding et al., 1997; From Bolender, 2010).

Figures 3.7 and 3.8 suggest the presence of a pecking order in biology, wherein one of the parts controls more connections than another. For example, in Figure 3.8 the dominance shifts from the dentate gyrus in the control to the presubiculum in alcoholism and Alzheimer's disease. By summarizing the contents of the organism code database (Figure 3.9), we can rank the dominant central organizers. As one might expect, the parts containing DNA (nuclei and mitochondria) exert the greatest control. However, the competition between the two controlling DNAs for resources adds another level of complexity when trying to diagnosis or predict changes in cell compartments (Bolender, 2017).

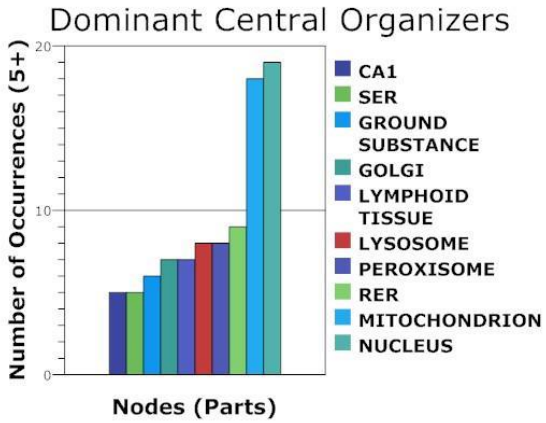


Figure 3.9 The dominant central organizers receiving the largest number of connections include the nucleus and mitochondrion. The tendency of cell organelles to key on specific parts may point to a first principle of cell design (From Bolender, 2010).

3.1.6 Parallel Complexities

A parallel complexity is defined as a collection of decimal ratios, which when expressed as alphanumeric strings, serve as a proxy for biology.

The parallel complexities currently being described at level 1 derive from databases containing morphological data. Since such databases come from two distinct sources (living and nonliving subjects), we can expect the two data sets to carry different methodological biases.

Such differences, however, can act to our advantage. MRI data collected from living patients tend to minimize methodological biases, whereas stereological data collected postmortem carry a much heavier load thereof. Understandably, part of the challenge in applying stereology to biology includes figuring out how to minimize the negative effects of such biases. MRI can help (Figure 3.19; Top).

3.1.7 Mathematical Mapping

Mathematical mapping consists of associating an element of one set with that of another, both of which share a common value. It serves as a quantitative measure of connectivity. Moreover, by depicting the relationship of parts to connections visually, we can begin to appreciate the extent and strategy of biological complexity.

We begin with the MRI data (volumes) collected from the brains of patients (Goldstein et al., 1999). Figure 3.10 represents a mathematical map of the cerebral cortex (control subjects), wherein 42 parts (blue dots) display a bewildering number of interconnections (red lines). The figure shows how biology can use it parts and connections to design the brain according to a set of well-defined rules. By mapping all 42 parts simultaneously, we begin with a global view of the human cerebral cortex as a complexity, which, in turn, can be unfolded to reveal local patterns and unsuspected relationships.

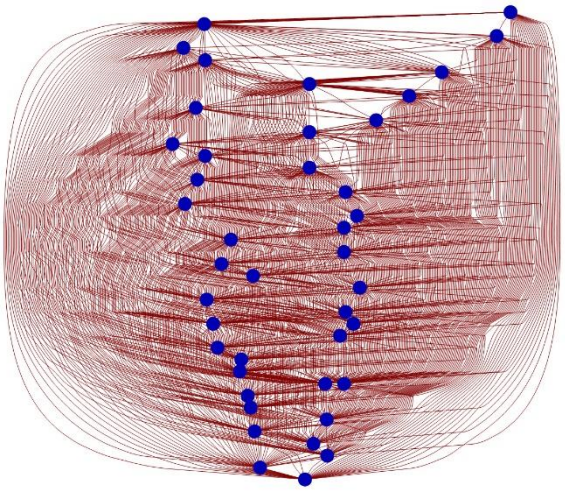
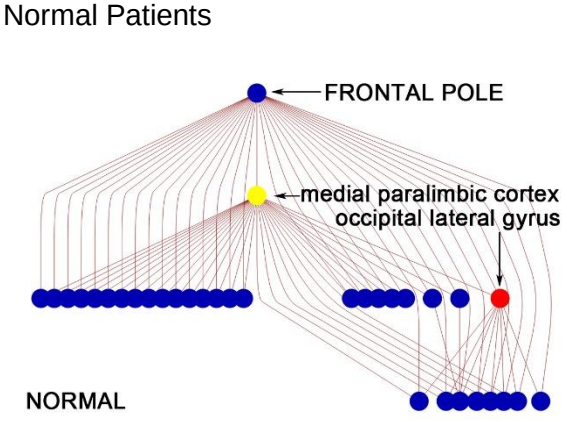


Figure 3.10 A mathematical map of the normal human cerebral cortex derives from forty-two parts displaying thousands of connections (Original data adapted from Goldstein et al., 1999; From Bolender, 2011).

3.1.8 Diagnostic Patterns

When we compare, for example, the frontal pole in normal individuals to those with schizophrenia, distinctly different patterns appear (Figure 3.11). In the original study, however, no such patterns were reported (Goldstein, et al., 1999). Notice that schizophrenia produces dramatic changes in the connectivity of the parts throughout the brain (Figures 3.11, 3.12). In such cases, the maps become diagnostic of the disorder.



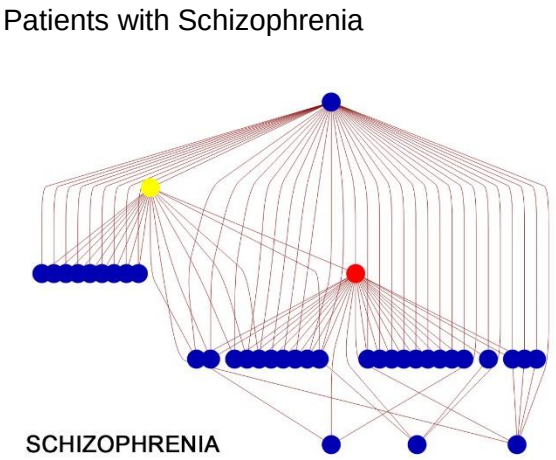


Figure 3.11 In schizophrenia, the relationship of parts to connections in the human frontal pole changes (Original data adapted from Goldstein et al., 1999; adapted from Bolender, 2011). Blue (at top of field) = frontal pole, yellow = medial paralimbic cortex; red = occipital lateral gyrus

Although complex patterns lend themselves to a graphical analysis, mathematical maps can also be condensed into equations. In Figure 3.12, for example, the equation for schizophrenia (dashed red line) derived from the cerebral cortex fits a polynomial equation ($y = 0.0485x^2 - 4.3726x + 98.06$ with an R^2 of 0.9854. Notice that it is readily distinguishable from the corresponding control curve ($y = 0.0315x^2 - 3.4193x + 87.57$; $R^2 = 0.9815$).

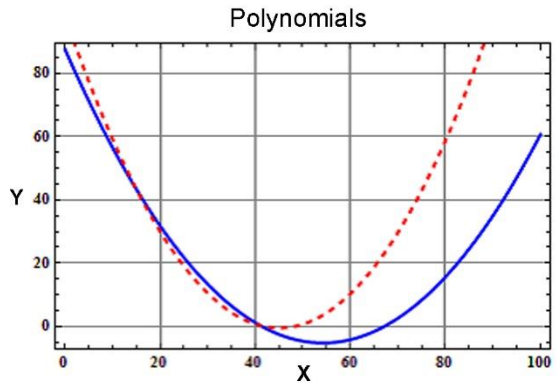


Figure 3.12 Diseases such as schizophrenia can be expressed as a polynomial equation (dashed red line), which can be distinguished from the one of normal patients (solid blue line) (Original data adapted from Goldstein et al., 1999; From Bolender, 2011).

3.1.9 Generalizing Disorders

The human brain displays a remarkable capacity for creating disorders. But what if these disorders are related and share many of the same roots and branches analogous to the familiar schemes of evolutionary trees? If, in fact, disorders are modular in design, then modular approaches to diagnosis and treatment might prove to be an effective clinical strategy.

Although a disorder of the human brain carries a distinctive set of mathematical markers, the same markers can appear in more than one disorder. This tell us that the complexity of a disease depends on multiple factors, including the composition of individual markers, the presence or absence of certain markers, and the total number of abnormal markers in play. In effect, we can use mathematical markers to characterize each disorder as a unique phenotype. This replicates phenotypically the approach currently being applied to genes.

Figure 3.13 includes a list of brain parts cross-correlated with disorders; it summaries the design strategy biology uses to create disorders. Notice that specific parts define a disorder, that different disorders often involve the same parts, and that relatively few parts (35/185 = 19%) account for most of the disorders. Schizophrenia with 26 changed parts and bipolar disorder with 20 produced the most damage with some parts being more vulnerable than others. The amygdala was involved in 13 disorders, caudate 13, hippocampus 10, putamen 10, and thalamus 9. [Note: Later in the chapter (Figures 3.31 and 3.32), a database based on counts of duplicate mathematical markers will put the hippocampus in first place.]

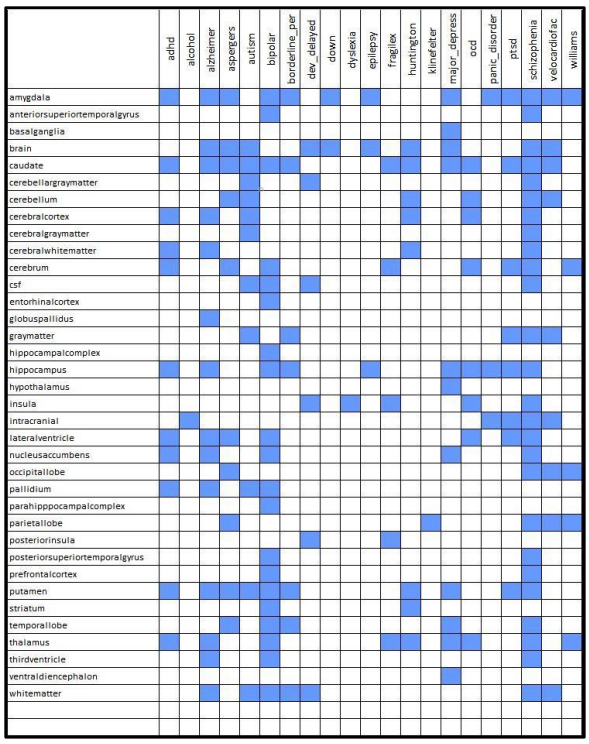


Figure 3.13 The figure summarizes the contribution of specific parts to 21 disorders of the human brain (summarized from 76 publications). Read the blue squares by row to identify the involvement of a given part in a disorder and the blue squares by column to identify the parts responsible for a given disorder. These parallel complexity data came from a diagnosis database consisting of triplets (Original data adapted from Kennedy, et al., 2012; From Bolender, 2012).

3.1.10 Creating Disorders with Modules

Figure 3-13 identifies disorders of the brain according to the parts involved. Once again, the patterns suggest that biology assembles things – new and old - from well-defined sets of parts – or modules - be they normal or abnormal.

By graphing 14 of the disorders listed in Figure 3.13, we can begin to see some of the details of this modular strategy. Schizophrenia, for example, represents the most extensive departure from the normal in that it carries at least 123 abnormal mathematical markers. By

adding the markers of 14 other disorders to the graph, we can see the relationship of schizophrenia to these disorders. Although schizophrenia remains the dominant player, it shares many of its parts (~30%) with the other disorders (Figure 3.14).

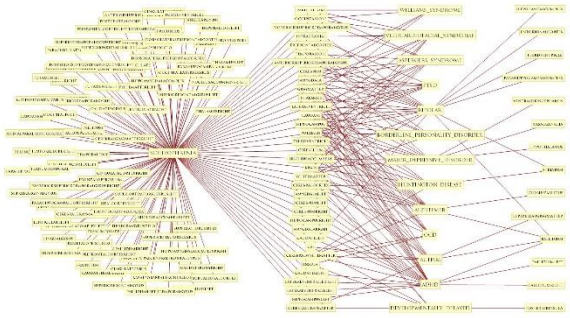


Figure 3.14 Disorders of the brain share many similar parts and connections configured as mathematical markers (Original data adapted from Kennedy et al., 2012; From Bolender, 2012). Enlarge the image to view details. [The figure is also available online.]

Notice in Figure 3.14 that the abnormal parts and connections of six disorders (Asperger's, borderline personality, OCD, PTSD, velocardiofacial, and William's) are the same as those found in schizophrenia. What could this mean? If the equivalent of a genetic script (or cohort) exists for schizophrenia, is it being read only in part to produce one of these six disorders or is schizophrenia a combination of many different disorders? What role might post translational processing be playing? These questions, of course, go to the heart of the disease process. The take away point here is that we can characterize these disorders objectively as quantitative patterns. Artificial intelligence (AI), for example, might find deeper patterns currently overlooked.

If we plot just three disorders (schizophrenia, bipolar disorder, and ADHD), the complex relationship of one disorder to another becomes easier to see (Figure 3.15). Notice that bipolar disorder and ADHD exist as distinct subsets of schizophrenia in that they share 80% of the same parts and connections. Likewise, a secondary relationship exists between the bipolar disorder and ADHD by sharing roughly 25% of the same parts and connections.

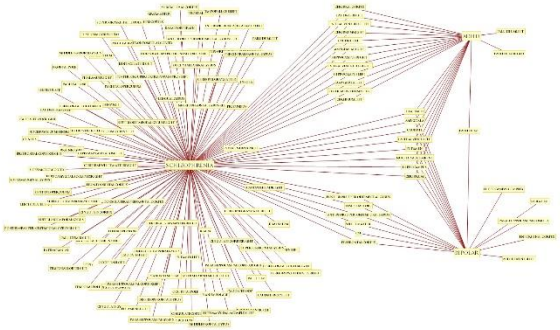


Figure 3.15 ADHD and bipolar disorder share many identical parts and connections (mathematical markers) with schizophrenia, as well as with each other (Original data adapted from Kennedy et al., 2012; From Bolender, 2012). Enlarge as needed.

Given what we have learned thus far, studying individual disorders within the context of all the disorders may turn out to be a more effective way of advancing our understanding. If, for example, we can induce an abnormal marker to revert to a normal one, then that solution might also apply to several other disorders. Mapping disorders with mathematical markers shows us where to look. What, for example, is the relationship between mathematical markers and clinical symptoms (emergent properties)?

3.1.11 Patterns and Problem Solving

Now that we can extract patterns from the biology literature, how do we go about using them to deliver solutions to problems of general interest? One way of answering the question is to identify problems that have stubbornly resisted solutions and then show how we can use patterns to find workable solutions.

How, for example, might we approach the problem of diagnosing disorders of the brain? The current approach, which relies on the analysis of subjective symptoms, suffers the

limitation of having to deal with disorders displaying many overlapping symptoms (Bolender, 2015). Let's try a fundamental shift in strategy. For our purposes here, this will include shifting the focus of a clinical diagnosis from subjective symptoms to objective measures.

Such a strategy, however, invariably leads to questions of a practical nature. Which database, for example, should we use - the MRI or the stereology? To be on the safe side, we'll try both.

This means that the problem we want to solve is one that involves collecting diagnostic data from two analogous but potentially dissimilar sources - living and postmortem brains. We want to answer two questions. Do both data sets yield the same results? If not, why not? In turn, the answers to these questions will serve to guide us toward meeting our larger goal of identifying objective diagnostic tests with improved reliability.

3.1.12 Diagnosing Schizophrenia

Since we have access to data coming from the same parts in living (Internet Brain Volume Database [IBVD]) and postmortem brains (Stereology Literature Database [SLD]), we can generate two diagnostic databases from the literature, both of which consist of triplet mathematical markers (ax:by:cz).

If we mix the mathematical markers coming from the two databases (MRI with 24 disorders and stereology with just schizophrenia), select the resulting duplicates (MRI = stereology), and plot them (Figure 3.16), the stereology markers should detect schizophrenia. Did they? No. What happened?

Figure 3.16 shows that the stereology markers missed the correct diagnosis by picking the bipolar disorder, instead of the correct one (schizophrenia), which displayed a distant fourth-place finish. Such a result tells us that many of the parts in living and postmortem brains may share the same names but not necessarily the same volumes and ratios.

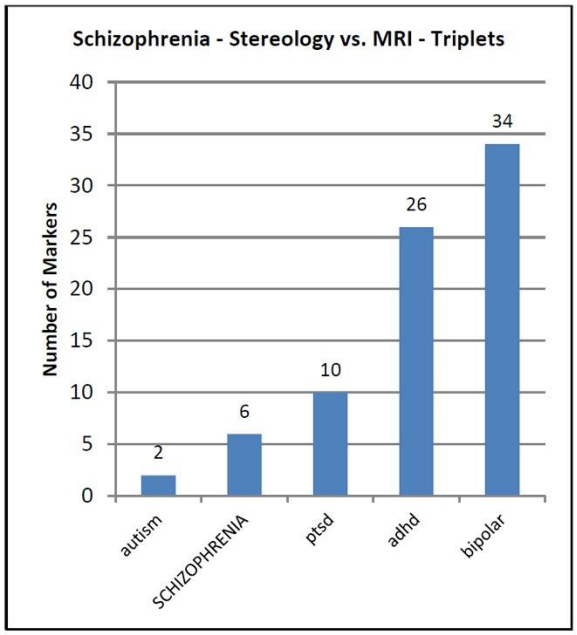


Figure 3.16 When mathematical markers for schizophrenia were taken from postmortem brains (stereology) and run against a panel of 24 disorders derived from living brains (MRI), the resulting duplicates led to an incorrect diagnosis (bipolar). The correct diagnosis – schizophrenia – ended up in fourth place. Such a result suggests that mathematical markers derived from living and non-living brains can be incompatible quantitatively. Notice that relatively few markers were in play (Original data adapted from Kennedy et al., 2012; From Bolender, 2013).

Why is such a finding, as shown in Figure 3.16, such a troubling result? Since stereological methods contribute to numerous studies of the brain, data based on potentially faulty reference volumes may be corrupting results, as well as hindering our efforts to reproduce experimental outcomes.

Is there a fix for this problem? Fortunately, yes. If we identify MRI data coming from living patients as the gold standard, then postmortem distortions can be corrected thereto (Bolender, 2013).

Although shrinkage corrections have been applied to the entire human brain, corrections at the level of individual brain parts have received little or no attention. We can mitigate this shortfall by introducing correction factors for a wide range of stereological estimates, as shown in Figure 3.17.

This figure uses volume data taken from the MRI and stereology databases to calculate ratios that identify the amount of swelling or shrinkage accompanying individual parts of the brain. Although the average shrinkage (11%) is consistent with an expected value of 10-15%, individual values paint a disturbing picture with volume distortions approaching 90%. As shown in Figure 3.17, the methodological biases include the swelling (>1) and shrinking (<1) of the named parts.

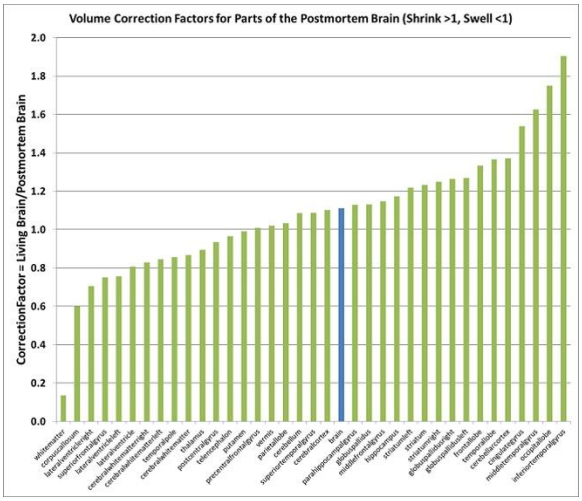


Figure 3.17 Volume correction factors for specific parts of the postmortem human brain show a wide range of values. A correction factor equal to 1 indicates no change, >1 shrinkage, and <1 swelling. The blue column identifies the brain, which needs a correction factor of 1.11 to account for shrinkage of about 11% (Original data adapted from Kennedy et al., 2012; From Bolender, 2013).

Armed with corrections for the volume distortions of individual parts (Figure 3.18), we can apply them to the stereological data used in our earlier attempt to diagnose schizophrenia with postmortem data (Figure 3.19).

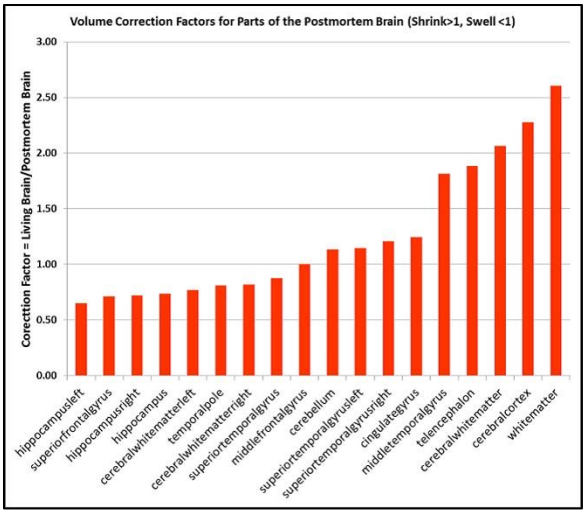


Figure 3.18 Volume corrections for the parts of the postmortem brains with schizophrenia exhibit a wide range of values (Original data adapted from Kennedy et al., 2012; From Bolender, 2013). They are used to correct the post mortem estimates to the MRI values, which serve as gold standards.

With the corrections applied (Figure 3.18), we can use the previously ineffective stereological data (Figure 3.16) to diagnose schizophrenia correctly (Figure 3.19; Top). Now, both data sets (living and postmortem) give the same diagnostic result.

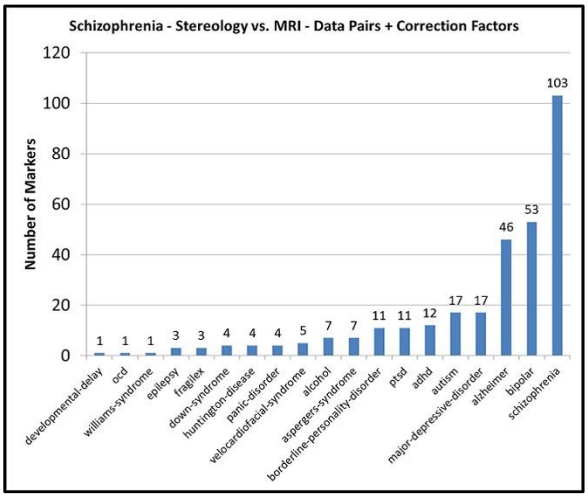


Figure 3.19 (Top)

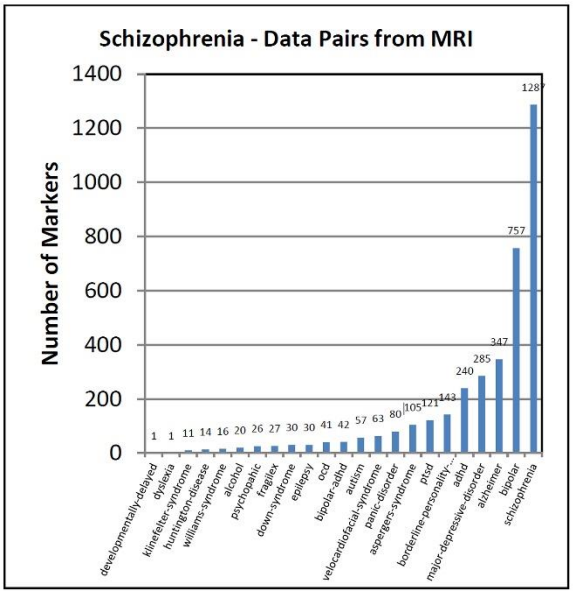


Figure 3.19 With correction factors applied to the distorted volumes, the stereological data deliver the correct diagnosis - schizophrenia. This puts the postmortem data of the brain back in play. Top: corrected stereological data; Bottom: MRI data (Original data adapted from Kennedy et al., 2012; From Bolender, 2013).

3.1.13 Diagnosing Disorders of the Brain

Trying to diagnose a disorder of the brain objectively - without knowing the diagnosis beforehand – can require a lot of work only to deliver an inconclusive result (Bolender, 2014). However, going from a known to an unknown becomes a better way to proceed because it's easier to spot and work around the many pitfalls hidden in such an undertaking. In practice, building a "bullet-proof" diagnostic database first, before testing it against unknown disorders, produces a far more satisfying result.

If we start with triplet markers (ax:by:cz) generated from 27 disorders, we can produce 277,039 mathematical markers from the data of 117 MRI papers stored in the IBVD. Alternatively, using quadruplet mathematical markers (ax:by:cz:dq) from 22 disorders, we can get 3,651,770 markers from 75 papers (IBVD).

Both diagnostic databases (triplet and quadruplet) can be designed to always deliver the correct diagnosis (Figures 3.20 and 3.21). Notice that both the number of papers (1-35) and mathematical markers (63 to 114,878) vary widely across the disorders.

TRIPLET DATABASE - UNIQUE MARKERS - 2014			
DISORDER	**MARKERS**	**PAPERS**	**DIAGNOSIS**
ADHD	27,499	6	YES
AFFECTIVE-PSYCHOSIS	494	2	YES
AGING	120	1	YES
ALCOHOL	984	2	YES
ALZHEIMER	28,568	5	YES
ASPERGERS	16,574	4	YES
AUTISM	2,921	11	YES
BIPOLAR	46,839	16	YES
BIPOLAR-ADHD	34	1	YES
BORDERLINE PERSONALITY DISORDER	2,847	2	YES
DEVELOPMENTAL-DELAY	948	2	YES
DOWN-SYNDROME	63	1	YES
DYSLEXIA	210	1	YES
EPILEPSY	276	2	YES
FRAGILEX	1,502	2	YES
HUNTINGTON-DISEASE	2,692	3	YES
INTRAUTERINE-GROWTH-RESTRICTION	676	1	YES
KLINEFELTER-SYNDROME	1,232	1	YES
MAJOR DEPRESSIVE-DISORDER	4,046	7	YES
OCD	3,091	1	YES
PANIC-DISORDER	1,858	2	YES
PRETERM	4,023	3	YES
PSYCHOPATHIC	938	1	YES
PTSD	5,288	2	YES
SCHIZOPHRENIA	114,878	35	YES
VELOCARDIOFACIAL	7,490	2	YES
WILLIAMS	948	1	YES
TRIPLET TOTALS	**277,039**	**117**	**100%**

Figure 3.20 When filtered to remove duplicates, a triplet database consisting of unique markers can diagnose a disorder – selected from a diagnosis database - correctly 100% of the time (Original data adapted from Kennedy et al., 2012; From Bolender, 2014).

QUADRUPLET DATABASE - UNIQUE MARKERS - 2014			
DISORDER	MARKERS	PAPERS	DIAGNOSIS
ADHD	240,286	3	YES
AFFECTIVE-PSYCHOSIS	1,008	1	YES
ALCOHOL	1,212	1	YES
ALZHEIMER	587,131	2	YES
ASPERGERS	338,064	3	YES
AUTISM	10,797	6	YES
BIPOLAR	771,734	11	YES
BIPOLAR-ADHD	450	1	YES
BORDERLINE PERSONALITY DISORDER	50,417	2	YES
DOWN-SYNDROME	541	1	YES
DYSLEXIA	63	1	YES
EPILEPSY	2,821	2	YES
FRAGILEX	6,372	1	YES
HUNTINGTON-DISEASE	22,410	2	YES
KLINEFELTER-SYNDROME	9,036	1	YES
MAJOR DEPRESSIVE-DISORDER	44,277	3	YES
OCD	38,685	1	YES
PANIC-DISORDER	14,982	1	YES
PRETERM	49,800	1	YES
PTSD	64,153	2	YES
SCHIZOPHRENIA	1,324,205	27	YES
VELOCARDIOFACIAL	73,326	2	YES
QUADRUPLET TOTALS	**3,651,770**	**75**	**100%**

Figure 3.21 When filtered to remove duplicates, a quadruplet database consisting of unique markers can diagnose a disorder – selected from the diagnosis database - correctly 100% of the time (Original data adapted from Kennedy et al., 2012; From Bolender, 2014).

Producing the results shown in Figures 3.20 and 3.21 proved to be a challenging task because the diagnostic process was continually being compromised by false positives and false negatives. The solution to this problem consisted of sequestering the diagnostic data in a data cage (Figure 3.22), wherein each disorder has its own set of unique mathematical markers. This guaranteed a correct diagnosis 100% of the time.

3.1.14 Diagnosing with a Data Cage

By encapsulating the diagnostic data in a cage (Figure 3.22), thousands to millions of unique mathematical markers can be vetted and shown to be reliable 100% of the time. The next step of the diagnostic process consists of constructing a visual interface that diagnoses a disorder by binding the markers of an unknown disorder to those of known markers.

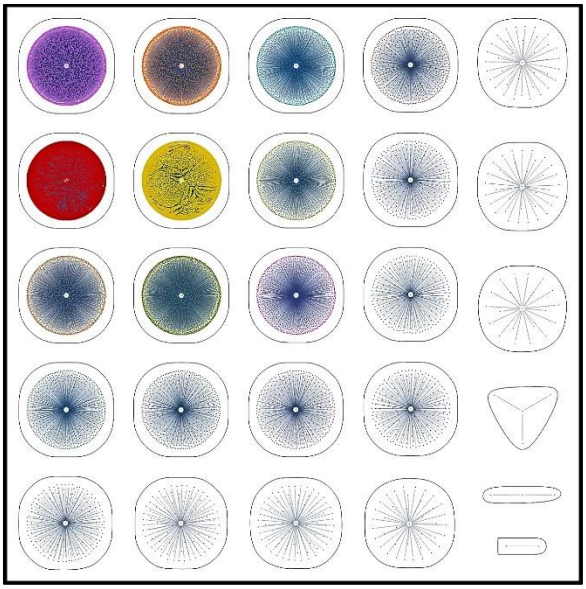


Figure 3.22 A data cage designed for the human brain includes a collection of 26 disorders – each of which is identified in the figure as a central point surrounded by points representing mathematical markers unique to that disorder (Original data adapted from Kennedy et al., 2012; From Bolender, 2014).

To show how the data cage works as a diagnostic tool, twenty markers of one disorder (bipolar) were copied, renamed as unknowns, added back to the cage, and plotted a second time (Figure 3.23). The diagnostic plot shows that all the "unknowns" attached to a single disorder – the one from which they originally came (bipolar). In principle, the data cage serves as a known standard to which unknown disorders can be compared. In effect, it becomes a simple, straightforward test.

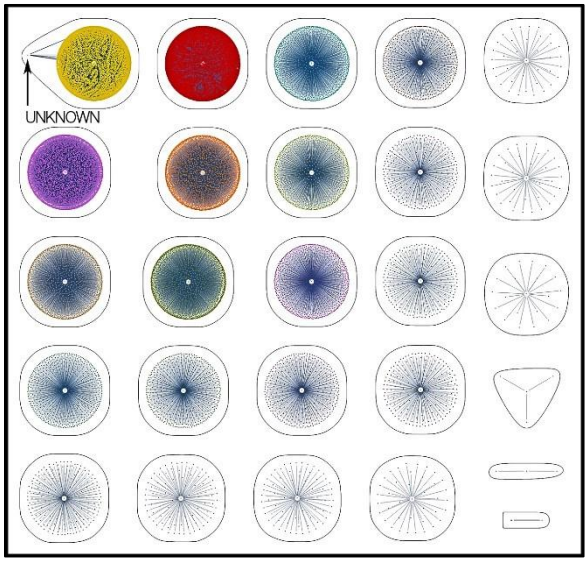


Figure 3.23 When an unknown disorder characterized by 20 mathematical markers is run against the entire set of 26 known markers, the unknowns (marked by an arrow) bind to the disorder sharing the same markers. The unknown markers were diagnosed correctly as belonging to patients having bipolar disorder. For purposes of illustration, the example was limited to 50,000 markers – out of 245,000 (Original data adapted from Kennedy et al., 2012; From Bolender, 2014).

Why was this objective approach to diagnosis successful? The current approach to diagnosing disorders of the brain requires advanced levels of clinical expertise confirmed by rigorous vetting. Mathematical markers captured and leveraged the clinical expertise contained within 172 publications, while, at the same time, collected many of the rules responsible for the disorders. The data cage proved to be successful as a diagnostic tool because it was based on unique identifiers (mathematical markers) that eliminated false positives and false negatives. By increasing the number of disorders and markers stored in the data cage database, we can expect it to work as a diagnostic tool with increasingly effectiveness in the general population.

A point worth noting is that data cages are likely to accommodate a wide range of diagnostic problems in clinical medicine. Since a diagnosis represents an exercise in recognizing and interpreting patterns, an objective approach has the advantage of putting millions of patterns in play, all of which also carry predictive properties. In effect, patient data – expressed as a complexity - come with built-in feedback loops, extending from the past, to the present and into the future. One can imagine a yearly checkup that begins with a head scan that characterizes one's current health and predicts probability-based outcomes accompanied by recommendations.

Note that mathematical markers with their ability to form universal biology databases would seem to be natural candidates for training and improving the effectiveness of deep and machine learning, both of which are key components of artificial intelligence (AI).

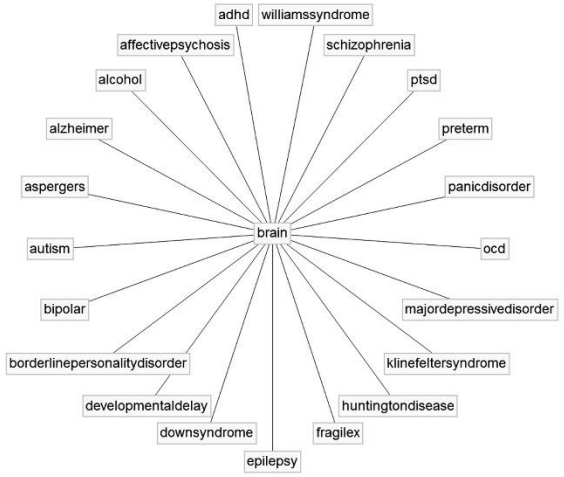
3.1.15 Exploring the Disease Process

Diagnosis, prediction, and unravelling the disease process all qualify as big data problems because of their complexity. Once populated with published data, however, these large clinical databases create new opportunities for discovery.

When translated into mathematical markers, for example, a diagnostic database derived from the IBVD can be used to explore patterns basic to the disease process. If, for example, a disorder is a function of a set of abnormal, but modular patterns, then a solution at the modular level for one disorder might apply to other disorders carrying similar modules. Since biological systems continually adapt and optimize within a rule-based framework, trying out new variations on a general theme may be biology's way of looking for new possibilities – new ways to adapt. When it goes too far, however, unintended consequences may occur.

3.1.16 Unfolding a Disordered Brain Objectively

For our purposes here, we will define a disordered brain as one containing 21 disorders - each characterized by a set of mathematical markers. Figure 3.24 identifies the disorders used for the unfolding.



cit_nu	mathematical marker	disorder
900	hippocampus1putamen1caudate1	alzheimer
631	hippocampus1putamen1caudate1	bipolar
456	hippocampus1putamen1caudate1	bipolar-disorder
456	hippocampus1putamen1caudate1	major-depressive-disorder
43	hippocampus1putamen1caudate1	schizophrenia
629	hippocampus1putamen1caudate1	schizophrenia

Figure 3.24 Top: A database of shared (duplicate) mathematical markers becomes a parallel complexity representing twenty-one disorders of the human brain (Original data adapted from Kennedy et al., 2012; From Bolender, 2015). Bottom: An example taken from the disordered brain database shows that the same mathematical marker can be shared by multiple disorders coming from different publications (cit_nu).

3.1.16.1 First Step

When applied to the disordered brain database, which contains disorders that can share mathematical markers (Figure 3.24), the CommunityGraphPlot (Mathematica, Wolfram Research, Inc.) locates five distinct clusters (Figure 3.25), four of which contain more than one disorder. The patterns displayed by the dark blue lines (connectivity) and the dots (mathematical markers and disorders) suggest that the subpopulations of the 21 disorders are inherently related because they share many of the same abnormal mathematical markers.

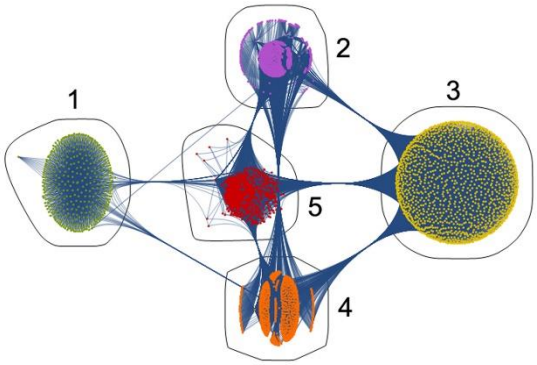
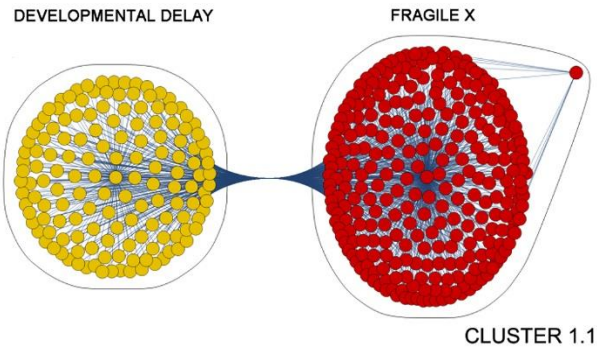


Figure 3.25 First Step: Shared mathematical markers (modules) from the aggregated abnormal brain (Figure 3.24) distribute – as communities - into five distinct clusters. Note that a dot can represent a disorder or a mathematical marker (which may or may not be a duplicate), whereas the dark blue lines represent connections (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

3.1.16.2 Second Step

The complexity of each cluster identified in Figure 3.25 unfolds to reveal the next grouping of complexity (Figure 3.26). Now the individual clusters labeled 1, 2, 4, and 5 in the First Step display clusters of their own. Cluster 3 – shown in Figure 3.25 - is the exception. It contains a single disorder (Alzheimer's disease).



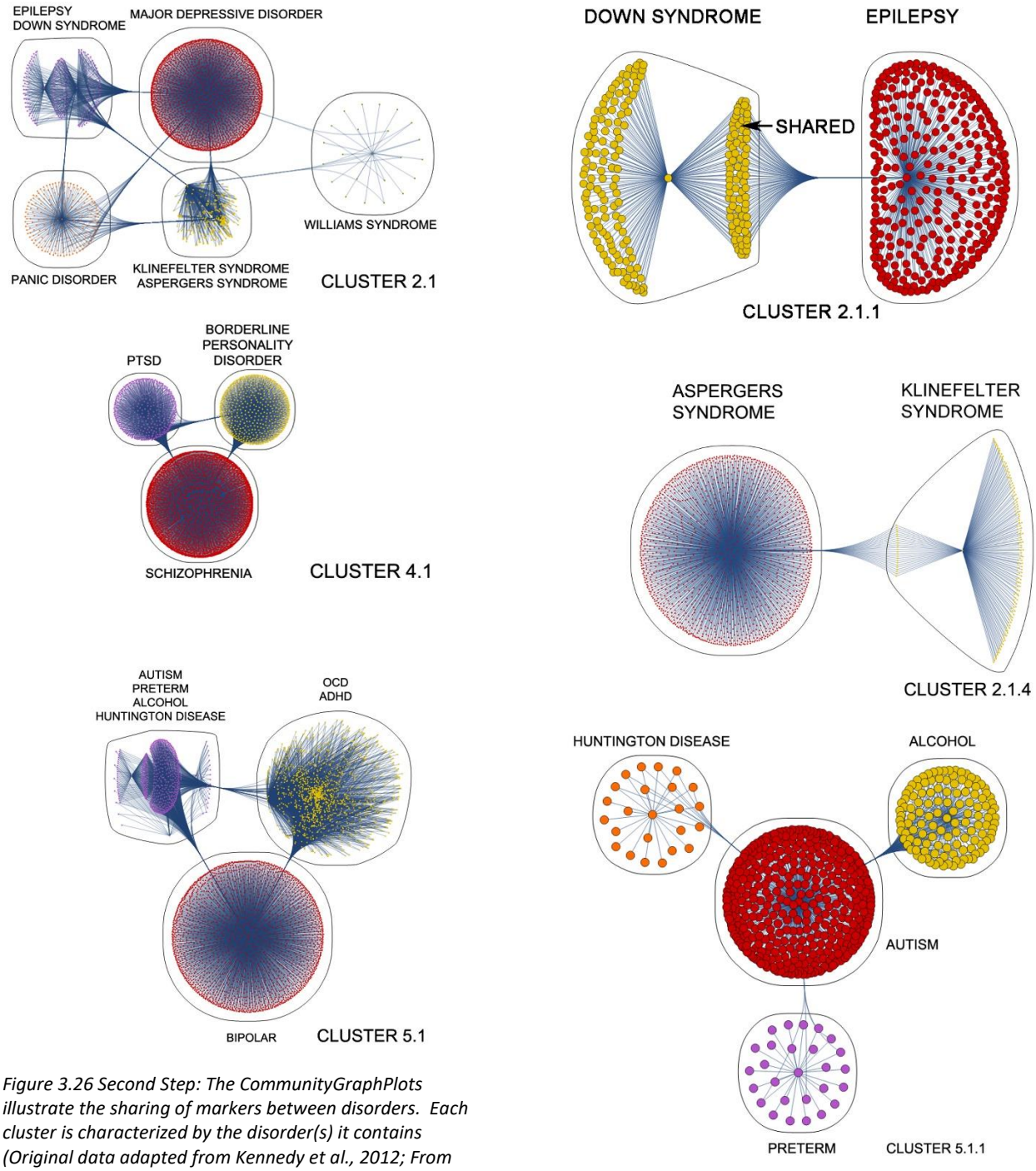


Figure 3.26 Second Step: The CommunityGraphPlots illustrate the sharing of markers between disorders. Each cluster is characterized by the disorder(s) it contains (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

3.1.16.3 Third Step

If a given cluster in Figure 3.26 carries more than two disorders, it is unfolded until it contains only two (Figures 3.27 and 3.28).

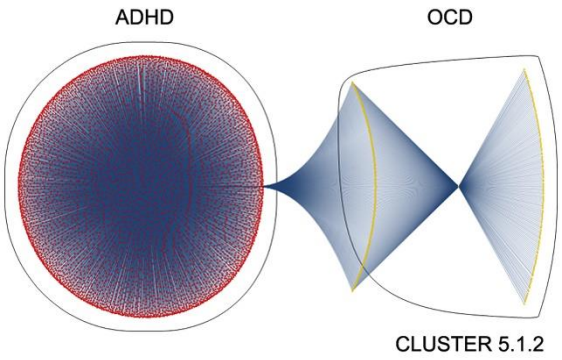
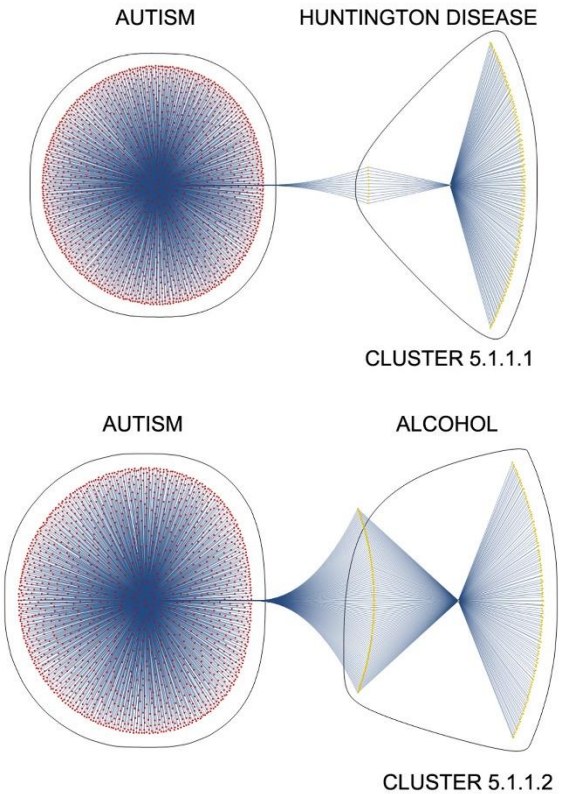


Figure 3.27 Third Step: Clusters containing more than two disorders in the Second Step are unfolded into clusters containing just two disorders. In such clusters, the shared markers appear as an intermediate, spindle shaped structure; see cluster 2.11 in Figures 3.26 and 3.30 (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

3.1.16.4 Forth Step

Cluster 5.1.1, for example, unfolds into three pairs of clusters (Figure 3.28).



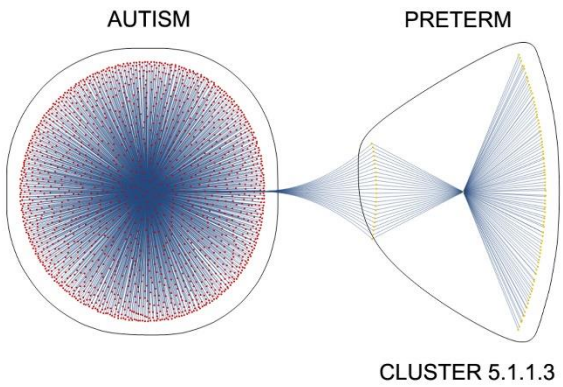


Figure 3.28 Forth Step: Cluster 5.1.1 is resolved into three clusters relating autism to Huntington disease, alcohol, and preterm (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

Figure 3.29 summarizes the disorders found in the five clusters identified in the First Step (Figure 3.25). Each cluster detects disorders most closely related – as defined by the sharing of mathematical markers. Keep in mind, however, that this summary reflects the contents of the current disordered brain database, which can be expected to change as more data are added (Recall Figure 3.20).

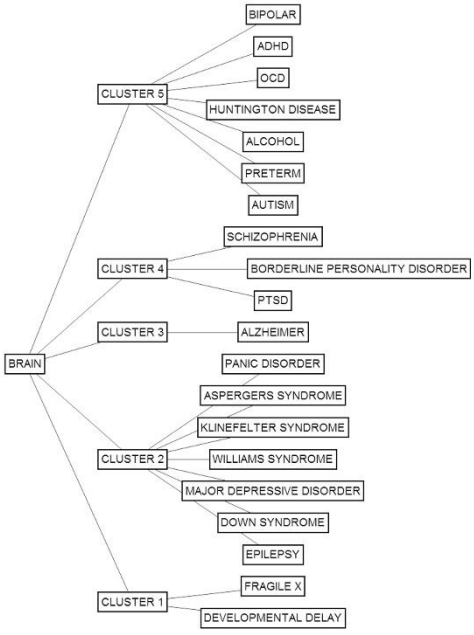


Figure 3.29 The graph shows the relationship of clusters to disorders in the aggregated human brain, based on their shared mathematical markers (From Bolender, 2015).

3.1.17 Sharing Markers

When reduced to a final cluster pair, we can identify the mathematical markers being shared by the paired disorders. Figure 3.30 details the relationship of ADHD to OCD by replacing the dots of Figure 3.27 with the alphanumeric strings of the mathematical markers. The overlap is striking in that the OCD cluster shares more than half (58%) of its markers with those of the ADHD cluster. Given such an extensive sharing of markers among disorders (Figure 3.25), it appears likely that substantial portions of many disorders are likely to map back to similar destinations in the genome or to somewhere else. If this turns out to be the case, identifying, targeting, and alleviating the most damaging disruptions may prove to be an effective strategy for managing groups of related disorders.

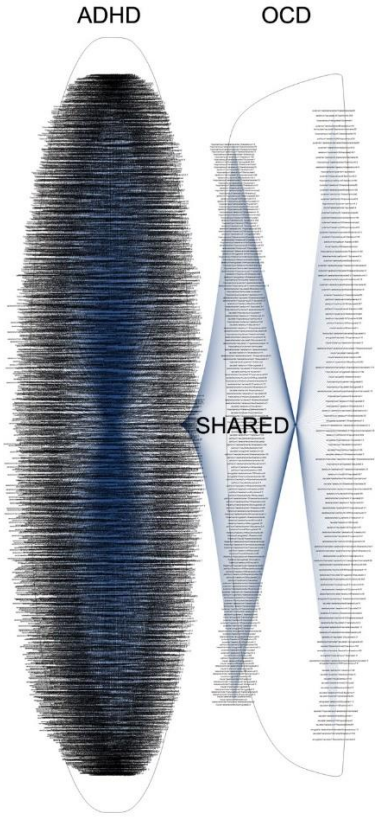


Figure 3.30 The clusters formed by ADHD and OCD show extensive sharing of mathematical markers (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

3.1.18 Key Players Producing Disorders

CommunityGraphPlots are useful in that they identify the preferences shown by disorders for specific markers, parts, and connections. Figure 3.31, based on a derivative of the IBVD (one of the disorders databases), shows that the hippocampus is the part of the brain most often involved in the disease process.

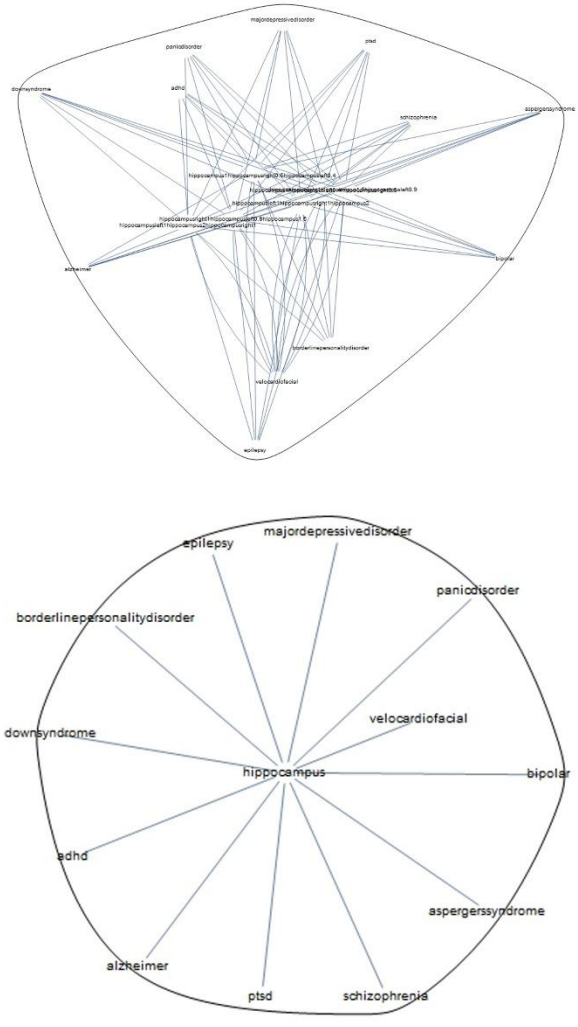


Figure 3.31 Duplicates >11. The hippocampus is the part of the brain contributing most often in the disease process (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

Figure 3.32 summarizes the relationship of parts to disorders according to the frequency of duplicate mathematical markers (>11 to >5).

Note that the Access database (Microsoft) was used to count the number of duplicate mathematical markers by applying filters. [Refer to the original report for details and worked examples (Bolender 2015)].

	>11	>9		>8			>7			>6			>5			
	HIPPOCAMPUS	HIPPOCAMPUS	AMYGDALA	HIPP-AMYG	PUTAMEN	CAUDATE	HIPP-AMYG	PUTAMEN	CAUDATE	HIPP-AMYG	PUTA-CAUD	TEMPLOBE	HIPP-AMYG-TEMPLO	PALLI-CEREBR	BRAIN-LATVEN	PUTA-CAUD-THAL
adhd	X		X			X		X						X		
alcohol					X			X								X
alzheimer	X		X	X			X								X	
aspergers-syndrome	X	X				X			X							X
autism			X		X			X								X
bipolar	X		X	X			X				X			X		
borderline-personality-disorder	X		X		X		X				X		X			
down-syndrome	X	X		X			X			X			X			
epilepsy	X		X	X			X			X			X			
huntington-disease																X
klinefelter-syndrome												X	X			
major-depressive-disorder	X	X			X		X					X				X
ocd														X		
panic-disorder	X	X		X			X					X	X			
ptsd	X	X				X			X		X				X	
schizophrenia	X	X		X				X		X	X				X	
velocardiofacial-syndrome	X		X			X			X				X			

Figure 3.32 The figure identifies the relationship of brain parts to disorders arranged left to right according to the descending numbers of duplicate mathematical markers (>11 to >5). These results suggest that a surprisingly small number of parts play a disproportionately large role in the disease process (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

The figure shows that most disorders of the brain depend importantly on abnormalities in just five parts – the hippocampus, amygdala, putamen, caudate, and temporal lobe – given the data set available at the time from the IBVD. It also summarizes the extensive sharing of parts that occurs across disorders. Taken together, these data continue to suggest a modular origin for the disorders of the brain.

CHAPTER 4

REPRODUCIBILITY

In recent years, a growing number of people and institutions have expressed concern over the inability of the sciences to duplicate studies and to demonstrate the validity of experimental results (Collins and Tabak, 2014; Begley and Ioannidis, 2015; Freedman et al., 2015; Roth and Cox, 2015; Engber, 2016). Moreover, metanalysis has shown that publications in the biomedical sciences are likely to be correct only about 20% to 30% of the time (Ioannidis, 2005; Colquhoun, 2014).

We now find ourselves deeply entrenched in what is being called a reproducibility crisis (Baker, 2016; Pellizzari et al., 2017). Since everyone seems to want less biological variation (to get better P values), the solutions being recommended focus largely on methodological issues.

For big problems, such as reproducibility, we always have two options - practical and theoretical. If the practical solution fails to get the job done, how might we approach the problem theoretically? We would have to derive reproducibility from first principles.

Since precision, accuracy, and reproducibility are all fundamental to the standing of a science, their absence – real or perceived – becomes reason for concern. This obviously creates a problem for the primer, a book professing to solve biology mathematically with data derived from the biology literature.

A central theme of the primer contends that interacting with biology as a complexity requires an approach well-suited to the task. It attributes the current shortcomings of biology as a science to the widespread belief that biology, chemistry, and physics can all operate successfully within the confines of a reductionist model. Given the reports of widespread absence of precision, accuracy, and reproducibility in the biological sciences, at least some of the fault must lie with a research strategy that reduces biology - a complexity - to a catalogue of isolated parts lacking in both connectivity and complexity. Since connections account for at least 50% of biology's complexity, why do we still insist on throwing them away?

The results presented in the primer stubbornly refuse to suffer a similar shortfall. By including both parts and connections when reconstituting biology's complexity, literature databases can now deliver highly reproducible patterns within and between species routinely. Given such an approach, the reproducibility crisis seen elsewhere has failed to appear here.

In this chapter, we explore reproducibility not as a statistical exercise, but instead as a byproduct of copying biology. But why not take the usual approach, which consists largely of trying to tamp down the variation coming from biology? Because such an approach is inconsistent with reality. Biology actively promotes individual variation as a form of adaptability essential to its survival. Moreover, biology prefers to get its reproducibility from well-established rules and first principles, which we are welcome to copy and use as our own measure of reproducibility.

4.1 Level 1 – Patterns

4.1.1 Knowing Where to Look

Biological data fall into two general groups, those with high variability and those with little or no variability. Isolated parts tend to exhibit high variability, whereas ratios and relationships of structure to function repeatedly display low variability. Biology uses both groups judiciously because to do so is advantageous to its short- and long-term interests.

4.1.2 Knowing What We Want to Do

Let's begin with reproducibility, which encapsulates concepts of both precision and accuracy. Two choices exist. To repeat an experiment, we can duplicate a previous result, or we can use literature databases to look for repeating patterns within and across species, experiments, and disorders. For the first choice, we might need a better statistical result (e.g., $P \leq 0.01$ to $P \leq 0.001$), but for the second we could collect the same patterns from many different settings and species. Both satisfy the notion of reproducibility. These options, however, are decidedly one-sided. A pattern based on parts and connections carries mathematical information with low variability, whereas biological parts have variability built into their design because such an arrangement supports adaptability. The statistical approach seems to work against biology, whereas the one based on ratios works with biology.

An example will help. The glycogen content of liver hepatocytes varies throughout the day in response to dietary intake and metabolic demands. Because of this built in variability, detecting a significant difference in the glycogen content between two different time points becomes understandably problematic. It requires that we synchronize events, which are occurring in control and experimental animals (Co = 5, Ex = 5), to within extremely close tolerances (e.g., $P \leq 0.01$). This built-in variability often frustrates our efforts. While the critics are quick to fault results with "poor" P values, it might make better sense to fault the premise of the experimental designs.

If biology incorporates variability into the design of its parts for good reason (e.g., adaptability), what do we gain by subjecting these parts to what may be unrealistic measures of precision, accuracy, and reproducibility? If an ability to demonstrate that significant differences exists between time points supersedes everything else, then the biology literature offers convincing evidence that we are failing to deliver research results routinely at higher levels of reliability ($P \leq 0.01$ to $P \leq 0.001$).

What's to be done? Change the rules. Shift the definition of reproducibility from simple to complex. Attach reproducibility to biology's rules, not to biology's need to allow individuals to respond to surrounding conditions.

If detecting and interpreting biological changes are the primary goals of an experiment, then designing such an experiment with well-defined patterns and rules instead of isolated data points should give us what we want, namely, equations with R^2s = 1 or ≈ 1. Moreover, to be more rigorous in the interpretation of our results, we can use databases to challenge the results across the literature rather than just by repeating the result of a single study. This gives us a far more robust and biologically friendly test for reproducibility.

4.2 Reproducibility Tests Applied to the Biology Literature

Let's begin by testing for the presence of reproducibility in the biology literature. We'll do this by looking for multiple copies of the same mathematical markers and connection ratios.

Figure 4.1 summarizes the volume, number, and surface area databases populated with mathematical markers attached to their connection ratios. The IBVD generated the volume data (MRI), whereas the cell counts (i.e., number of nuclei) and surface area data came from the stereology literature database (derived from light and electron microscopy).

Notice in Figure 4.1 that the MRI data detected the largest percentage of duplicate mathematical markers (41.6%), whereas cell counts (N) the least (1.1%). At 5.7%, surface areas displayed an efficiency rating five times greater than that for the cell counts (number) – despite its smaller sample size. The order of magnitude difference in the number of duplicate markers coming from the MRI and stereology data sets calls attention to the differences that seem to exist between data coming from living and nonliving sources.

Data Types →	**Volume**	**Number**	**Surface**
Data Sources→	**MRI**	**LM**	**TEM**
Mathematical Markers	374,906	108,824	21,402
Duplicates (≥3)	155,891	1,236	1,221
Efficiency (Reproducibility)	41.6%	1.1%	5.7%

Figure 4.1 The table summarizes the sample sizes used for the reproducibility tests. Abbreviations include MRI (magnetic resonance imaging), LM (light microscopy), and TEM (transmission electron microscopy); (From Bolender, 2016a).

For our purposes here, global counts of duplicates (≥3) will serve as the measure of reproducibility.

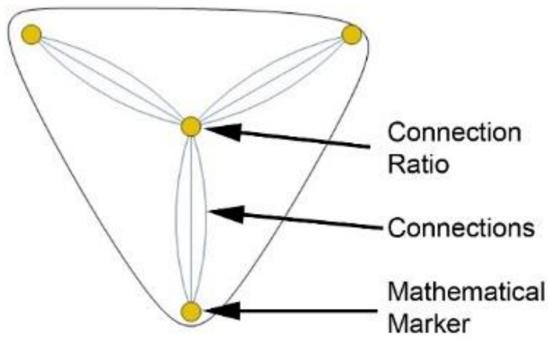
4.2.1 Defining the Reproducibility Test

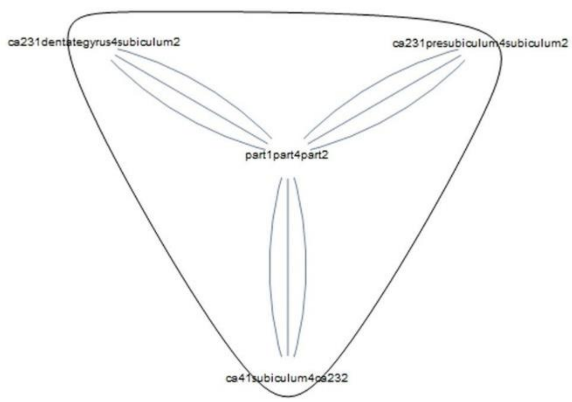
The test uses the CommunityGraphPlot of Mathematica to display the relationship of mathematical markers to their connection ratios (Figure 4.2). Units of complexity - identified here as rosettes – characterize reproducibility in the literature as a set of global patterns, whereas a single unit serves to quantify the duplication of mathematical markers within a given rosette (the number of blue lines in the "petals"). Both serve as global measures of reproducibility. In each unit, the number of duplicate markers is equal to the number of lines connecting a mathematical marker to its connection ratio.

For the MRI volume data, the connection ratio derives from three numerical values each representing the volume of a different part (V_{part1}:V_{part2}:V_{part3}). In Mathematica, the plots relate mathematical markers (a1b0.05c3) to their connection ratios (part1part0.05part3), as shown below in the example below:

"a1b0.05c3" → "part1part0.05part3".

The strength of the test depends on the number of connecting lines (forming blue spindles) and the total number of units (rosettes) in play. The number of different mathematical markers associated with a given connection ratio is a measure of the preference given to that ratio by biology. To pass the test, reproducibility must exist within and across many such units.





mathematical marker		connection ratio
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",
" caudate1caudateleft0.4hippocampusright0.4	"->"	part1part0.4part0.4 ",

Figure 4.2 Reproducibility can be approached quantitatively using units of complexity that display specific repeating patterns. By plotting mathematical markers (peripheral dots) against their connection ratios (central dot), we can test for the presence of reproducibility by counting the number of lines connecting the dots (3 in this case). Moreover, the plot identifies the number of different mathematical markers using the same connection ratio (3 shown). Since only one copy of a given marker is taken per paper (or experiment), 3 or more (≥3) connections signal the presence of reproducibility – at the global level (From Bolender, 2016a). Note that the Excel table shows the data set that produced one "petal" of the rosette shown in Figure 4.3. Each row of data came from a different source (paper or experiment).

The design of the test makes it quite difficult to pass. Triplet mathematical markers consisting of six variables must appear as duplicates in at least three different papers - or separate experiments - to qualify. Even so, the databases demonstrated repeatedly that the biology literature meets and exceeds this minimum requirement routinely (see Figures 4.3, 4.4, and 4.5).

4.2.2 Testing MRI Volumes

Since biological complexity remains intact in living subjects, the test was applied first to MRI data coming from patients. The MRI database used to test for reproducibility was derived from the IBVD.

Consider, for example, the rosette shown in Figure 4.3 (taken from Figure 4.4). Such an image could not appear unless biology and the biology literature were taking and passing the same test. The rosette shows that the same mathematical marker exists in as many as 13 of the 120 papers populating the IBVD, even when only a small fraction of the total markers are used for the plot (2,500 out of 155,891).

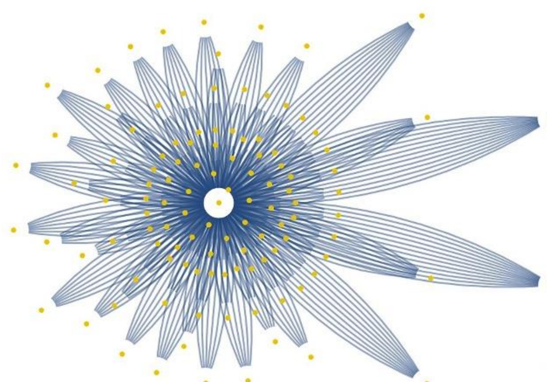


Figure 4.3 The pattern illustrates biology's approach to reproducibility. When MRI triplet data are plotted against a connection ratio, the mathematical markers form a rosette (one shown above) centered on a unique connection ratio. Multiple blue lines, which make up the spindle-shaped "petals," show that the same mathematical marker (peripheral yellow point) is being duplicated by multiple papers in the database (Original data adapted from Kennedy et al., 2012; From Bolender, 2016a).

Figure 4.4 illustrates the collection of rosettes produced by a small sample of the mathematical markers collected from the literature (IBVD), which consisted of 120 publications. The image shows the extent to which identical mathematical markers can exist - with a high degree of reproducibility - across publications and conditions.

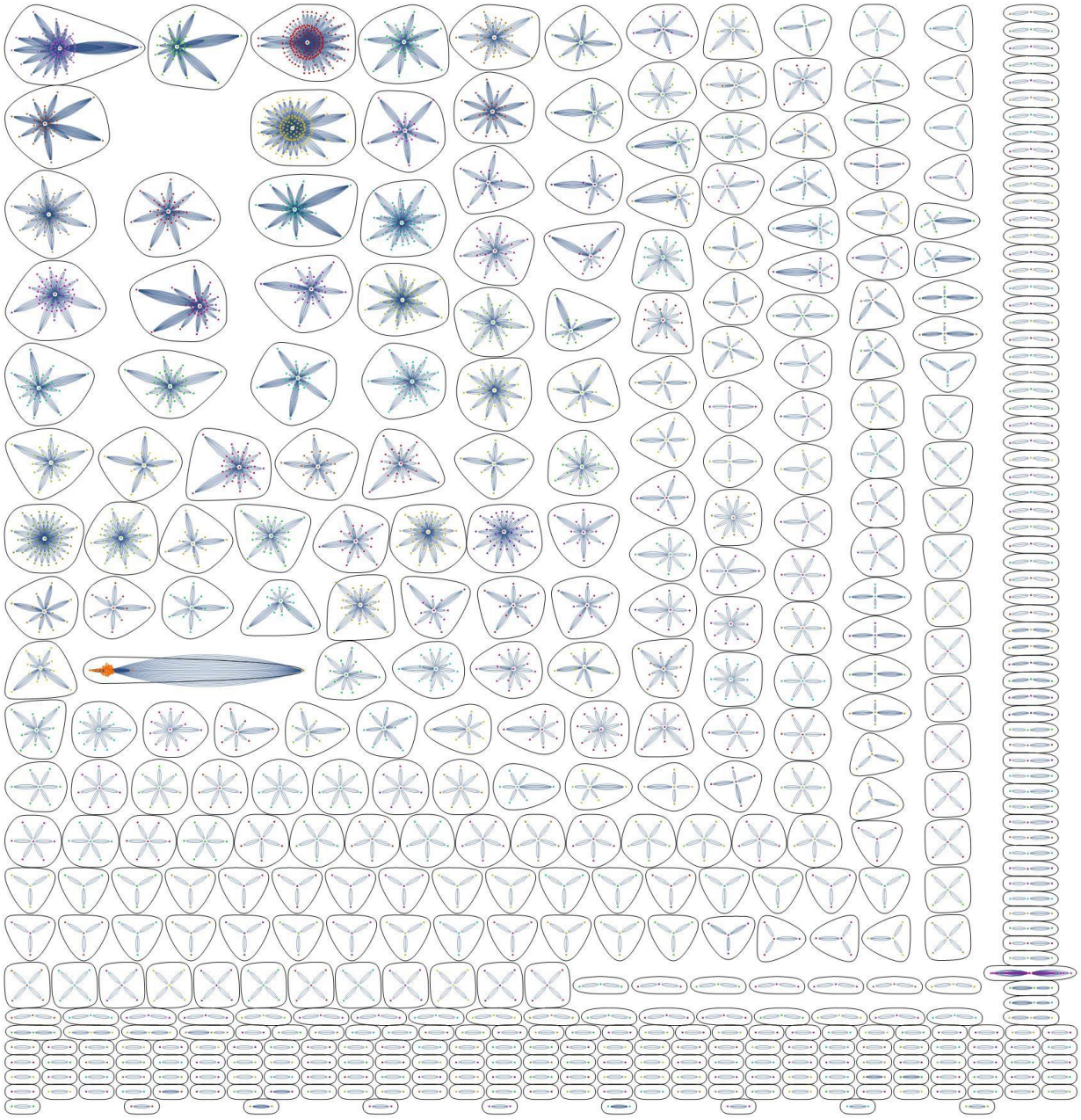


Figure 4.4 The MRI volume data of patients produce a wealth of global data as indicated by the many units (rosettes) and the frequency of duplicates (seen as dense blue spindles) linking mathematical markers to their connection ratios. Note that the figure uses just 2,500 of the 155,891 duplicate markers (≥3). The plot shows that reproducibility plays a key role in biology and exists abundantly in the biology literature (Original data adapted from Kennedy et al., 2012; From Bolender, 2016a).

Since reproducibility is the expression of biology's ability to maintain the ratios of its parts accurately, assembling a reproducibility test for the literature requires little more than copying biology's rules of design. Mathematical markers evidence this ability by displaying multiple copies of the same patterns distributed across the literature, wherein accuracy, precision, and reproducibility go hand in hand.

4.2.3 Testing Stereological Surface Areas

Now let's see what we can learn from the postmortem data of stereology. When we plot mathematical markers against their connection ratios, the surface area data also pass the reproducibility test (Figure 4.5) – more convincingly than the one for cell counts (numbers), but less convincingly when compared to the MRI data (Figure 4.4). Such a result makes a compelling argument for using data derived from living sources as our gold standard – based on the results presented.

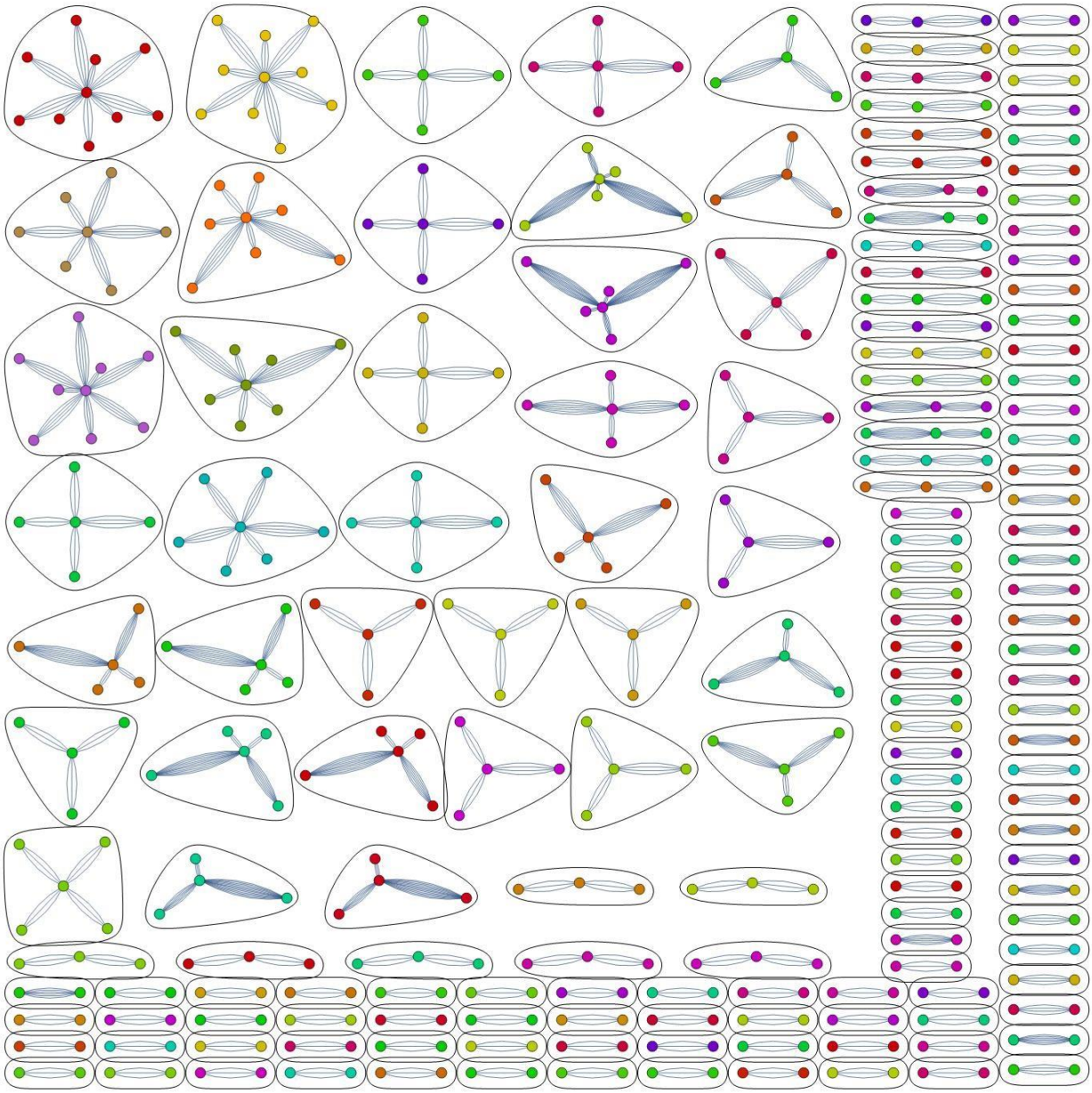


Figure 4.5 The surface area data (derived from cell organelles) passed the reproducibility test (many rosettes with dark blue spindles are displayed); (From Bolender, 2016a).

4.2.4 Testing the Stereology Literature

To biology, reproducibility would seem to include the patterns it expects to see phenotypically when the same genes are expressed in similar or different species - under similar conditions. If true, then reproducibility should exist throughout the biology literature. Let's test this idea.

If we plot mathematical markers for the surface areas of cell organelles against their citation numbers (displayed in boxes; Figure 4.6), we can generate a global view of reproducibility as it exists in the stereology literature database. Since the figure shows that duplicate mathematical markers exist routinely across the literature, we can argue that reproducibility exists as a property mutually shared by biology and the biology literature. This is reassuring, even though the postmortem data of stereology have a decidedly lower efficiency score than the one accompanying the MRI data (Figure 4.1). Recall that postmortem data, which carry a heavier load of bias, tend to be noisier than data coming from living sources. Bear in mind, however, that the stereological database includes a far more heterogeneous data set than that of the MRI database.

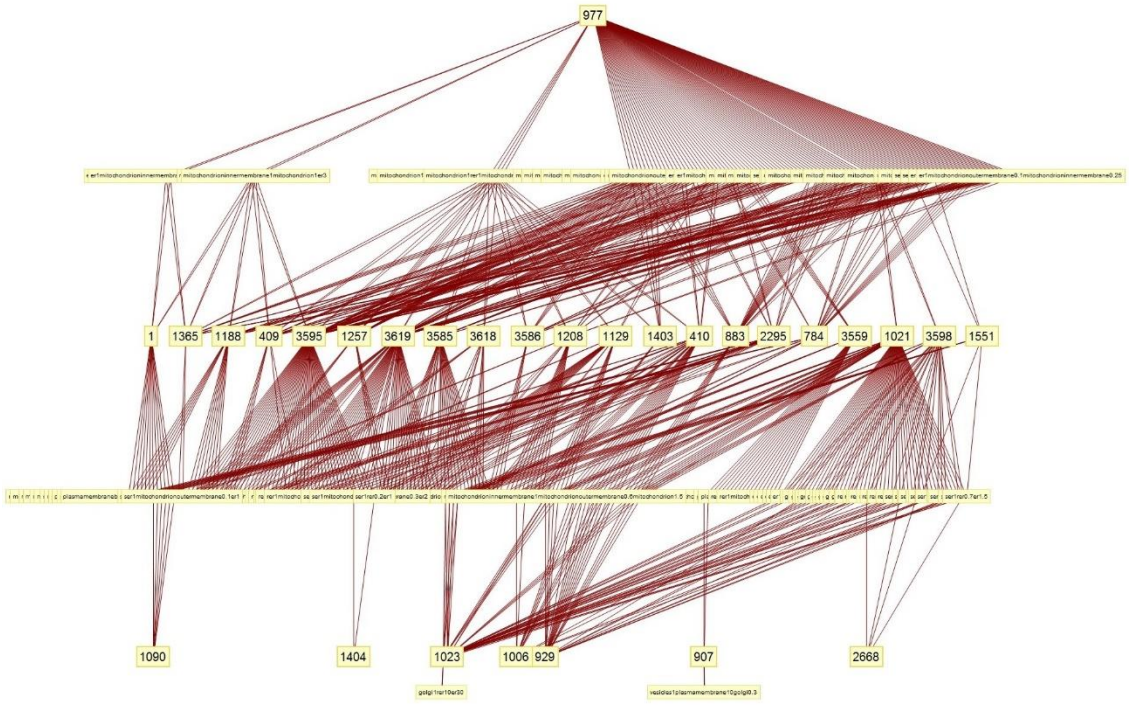


Figure 4.6 When surface areas of organelles coming from a wide variety of cell types and species (frogs to humans) are translated into mathematical markers and related to their citation numbers (boxed), global patterns appear across the stereology literature. Such patterns indicate that global reproducibility exists in published data (From Bolender, 2016a).

4.3 Reproducibility as a Complexity

By treating reproducibility as a complexity made up of parts and connections (Figure 4.2), it becomes easier to demonstrate and verify. Reproducibility exists as a quantitative property of biology, detected here as a core ratio being

shared by a well-defined set of mathematical markers within and across species.

A reproducibility test run under complexity rules offers four advantages: it (1) uses both parts and connections, (2) can compare and verify the results of one study against many, (3) works for most data types, and (4) applies the same rules locally (one paper) and globally (many papers).

While it seems likely that DNA codes - directly and indirectly - for both parts and connections, we don't know if they are being controlled separately or together. If biology controls them separately, then disorders – or for that matter any type of phenotypic change – is subject to at least two distinct levels of oversight. Disorders of the brain, for example, could be explained by abnormal parts, abnormal connections, or some combination of the two. A reproducibility test based on mathematical markers accounts for all three possibilities.

4.3.1 Unfolding Reproducibility

Reproducibility behaves like a complexity when approached as a collection of parts, connections, concentration ratios, and mathematical markers. If, for example, we look at just the connection ratios (recall Figure 4.2), an otherwise heterogeneous data set consisting of volumes, surfaces, and numbers of parts suddenly displays a high degree of connectivity (Figure 4.7). This tells us that duplication – used here as our measure of reproducibility – occurs at at least two levels of detail (mathematical markers and connection ratios).

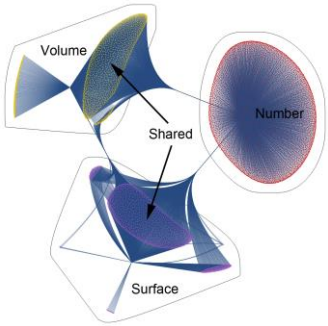


Figure 4.7 Although the volumes, surfaces, and numbers of parts represent distinctly different data types, their connection ratios show widespread similarities. In effect, different data types can share the same rules – at the level of connection ratios (From Bolender, 2016a). This and similar plots use the CommunityGraphPlot (Mathematica) to identify clusters of related patterns.

If, in turn, we plot the connection ratios of Figure 4.7 against species (Figure 4.8), the animals become grouped according to the similarity of their connection ratios. Compared to mathematical markers, however, the connection ratios detect reproducible patterns far more effectively by removing the specificity attached to the names of mathematical markers. They are uniquely suited to the task of detecting biological patterns at a level of detail below that of mathematical markers. In both cases, however, we are looking at the phenotypic projections of genetic control mechanisms.

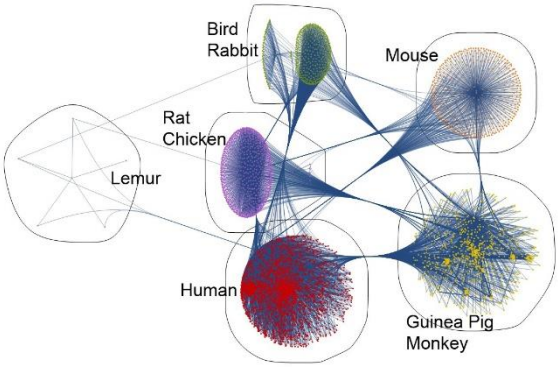


Figure 4.8 Even after the connection ratios of the volume, surface, and number databases were combined and plotted against species, the connectivity persisted. Note that these data include parts that range in size from organs to organelles, respond to a wide range of experimental conditions, and carry different

methodological biases. Once again, we see reproducibility behaving as a key unifying principle in biological systems (From Bolender, 2016a).

4.3.2 Reproducibility in the Liver

Organs consist of cells, which we tend to study one cell type at a time. However, the properties associated with an organ depend on the interrelationships of all the cells and subcompartments contained therein. We can view these inter- and intracellular relationships as global patterns by detecting them with mathematical markers and connection ratios.

Figure 4.9 uses the mathematical markers of hepatocytes, fat storing cells, endothelial cells, and Kupffer cells of the rat liver to look for reproducible patterns in the design of these cells. When we plot the mathematical markers for the surface areas of cell organelles against the names of the individual cell types (e.g., "er1rer0.5ser0.4" → "hepatocyte"), the results indicate that all four cell types can share identical markers.

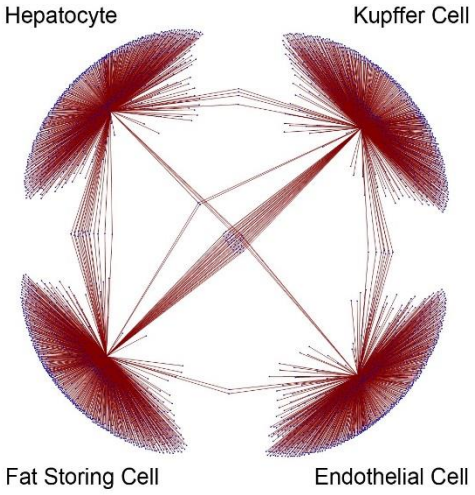


Figure 4.9 In the rat liver (Adapted from Blouin et al., 1977), hepatocytes, fat-storing cells, Kupffer cells, and endothelial cells display unique and shared mathematical markers. Notice that all the cells are interconnected, which suggests that they are expressing the same genes in the same way (From Bolender, 2016a). Such a finding is consistent with cells based on a modular design.

If, however, we replace the mathematical markers of Figure 4.9 with their connection ratios ("part1part0.5part0.4" → "hepatocyte") in Figure 4.10, the number of cell-to-cell connections increases strikingly. This tells us – once again - that cells populating an organ carry two distinct and reproducible patterns based on ratios, one with named parts (mathematical markers) and the other without named parts (connection ratios).

Such patterns based on quantitative ratios suggest that we can detect phenotypically – at multiple levels - the genetic programming of cell organelles by the presence (or absence) of mathematical markers and connection ratios. Such an arrangement allows biology to compile phenotypic recipes by changing parts and connections. This information may prove helpful in describing and explaining complex changes associated with differentiation, growth, and change – both normal and abnormal.

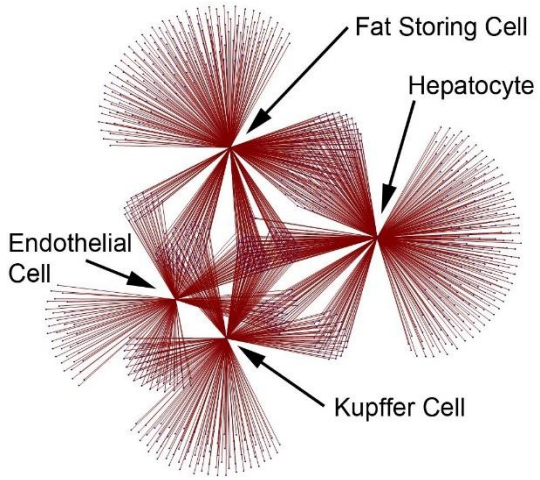


Figure 4.10 When the mathematical markers of Figure 4.9 are replaced by their connection ratios, the connectivity between the cells increases markedly. Now the cells appear tightly interconnected phenotypically (From Bolender, 2016a).

4.3.3 Reproducibility in the Literature

Notice in Figure 4.10 that we can identify two distinct populations of connection ratios, one unique to each cell (unshared) and another pool of ratios being shared by the four cell types. Why is this the case? Do the shared ratios signal a coordinated control of the cell types or are they just house-keeping ratios – common to cells in general? Will, for example, the shared ratios change when the hepatocytes are tasked with a specific job, such as metabolizing a drug or toxin? In other words, will a local change in one cell type (hepatocytes) influence its neighboring cells? The point in advancing such questions is that we can now ask and answer more probing questions using graphical analysis.

To visualize the pattern of cell to cell connectivity at a global level, we can plot the organelle surfaces of all the cell types in the stereology literature database against their mathematical markers (Figure 4.11). The pattern of shared markers appears once again. Of course, this is the pattern we would expect to see because we know that many species share many of the same genes.

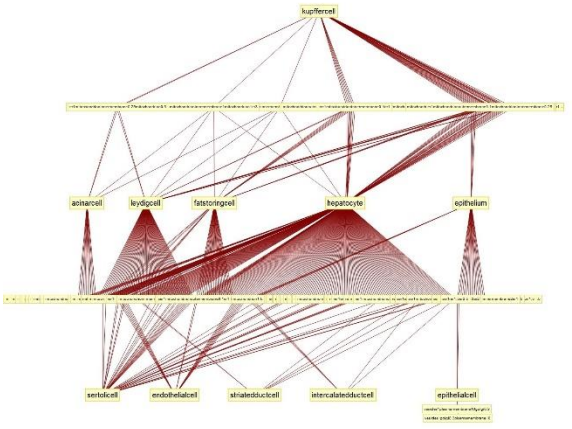


Figure 4.11 Widespread sharing of mathematical markers – derived from the surface areas of cell organelles - occurs within and across species (From Bolender, 2016a).

If we take the same set of mathematical markers shown in Figure 4.11 and substitute the citation numbers for the cell types, we get a global view of cell to cell reproducibility as it exists in the stereology literature database (Figure 4.12).

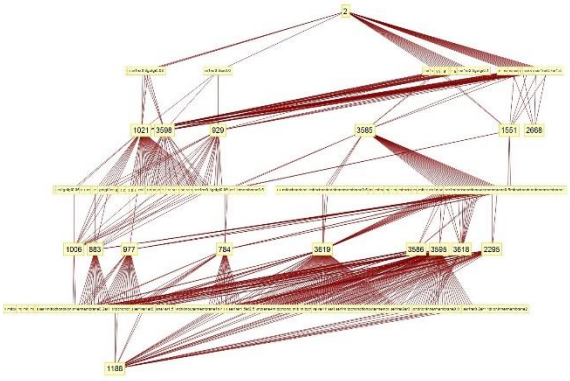


Figure 4.12 The plot shows reproducibility as a global property of published data. Recall that the mathematical markers being used here include alphanumeric strings consisting of six variables – making it a hard test to pass (From Bolender, 2016a).

When we replace the mathematical markers of Figure 4.12 with their connection ratios (Figure 4.13), the resulting view of the literature suggests an underlying global pattern of even greater reproducibility.

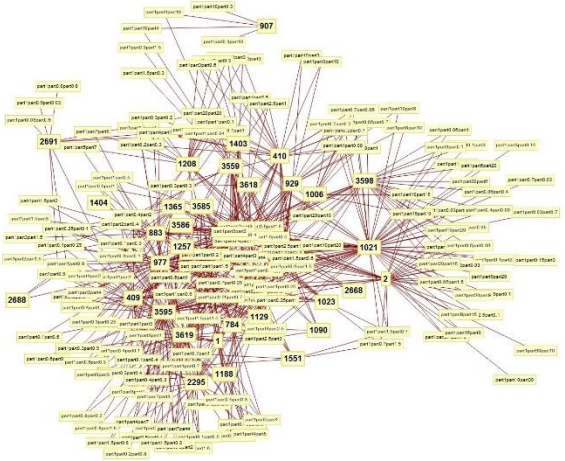


Figure 4.13 The plot displays the relationship between citations and connection ratios, as derived from the surface areas of cell oranelles. It illustrates the presence of widespread reproducibility within the stereology literature, at the level of connecion ratios (From Bolender, 2016a).

If we liken the presence of global data to reproducibility, it now seems likely that the biology literature contains far more reproducible results than previously imagined. Indeed, demonstrating reproducibility seems to be a function of knowing how and where to look for it.

4.3.4 Reproducibility in the Changing Brain

When using complex data types, reproducibility seems to exist everywhere we look. With MRI data (IBVD), for example, we can look for duplicate patterns by calculating and plotting the connection ratios of normal patients (C) against those diagnosed with 24 different disorders of the brain (E):

"C" → "part1part20part1.5" ,

"E" → "part1part1part0.6".

Figure 4.14 shows two sets of connection ratios (C, E) separated by a shared set wherein E=C. [Note that duplicate copies of the markers (E=E and C=E) were removed from each paper before aggregating the data and plotting the connection ratios.] Although the central, shared category (C=E) shows that many of the connection ratios remain normal, many more become abnormal (E). Since the figure identifies distinct populations of connection ratios for normal (C) and abnormal (E) brains, the connection ratios might also support diagnostic models similar to those based on mathematical markers (Figure 3.20), but at a coarser level of sensitivity. Figure 4.14, however, leaves fundamental questions unanswered. From where do the abnormal connection ratios come? Do the abnormal ratios map back to damaged genes, to errors in post translational processing, or to something else?

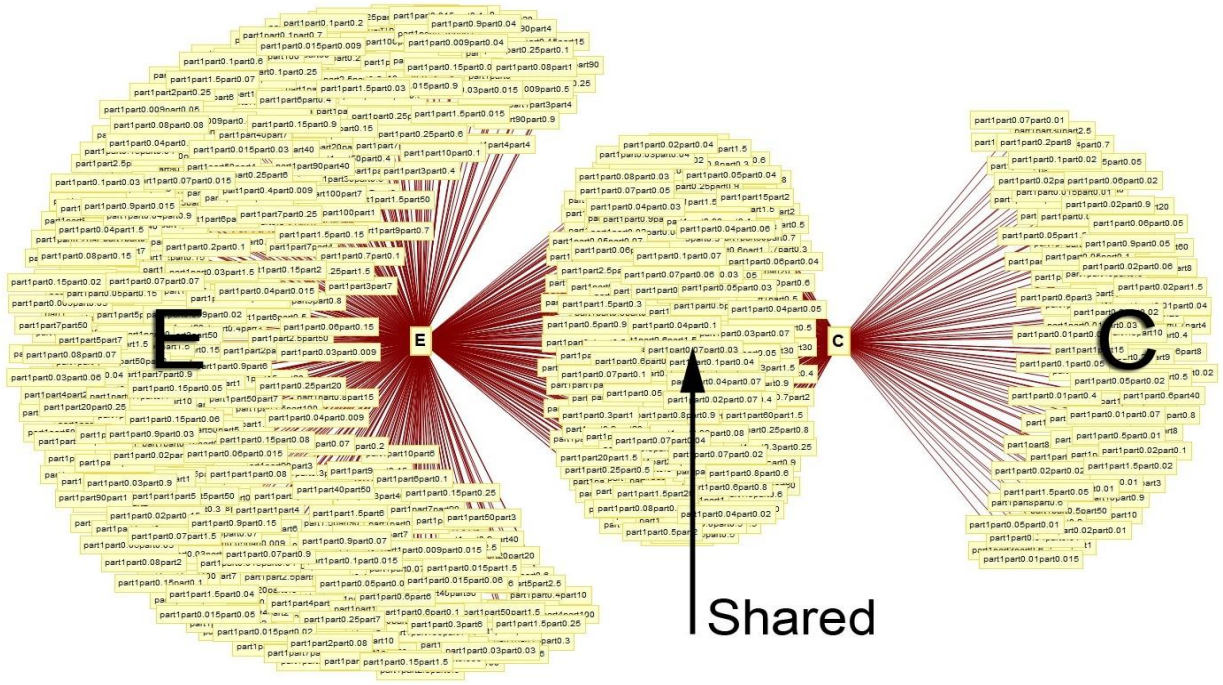


Figure 4.14 When translated into connection ratios and plotted, the MRI data of the IBVD form three distinct groups of connection ratios – abnormal (E), shared (E=C), and normal (C). The plot suggests that disorders of the brain result from widespread changes *(Original data adapted from Kennedy et al., 2012; From Bolender, 2016a)*.

4.3.5 Reproducibility Trap

Recall the standard definition of reproducibility:

Standard Definition: *"Reproducibility is defined as an ability to duplicate the results of an experiment either by the same researcher or by an independent one."*

Let's replace the standard definition with one more consistent with the way biology handles reproducibility.

New Definition: *Reproducibility is defined as an ability to duplicate a biological complexity with little or no variation.*

Consider the human brain. If, for example, we select estimates for the amygdala from the IBVD and plot them, we find a widely dispersed array of data points (Figure 4.15), illustrating the typical biological variation we have come to expect. If we begin by assuming that all the estimates are correct, which is the only fair thing to do, then the standard definition becomes open to criticism. By assuming that reproducibility can be tested locally (when a global test is indicated), it deliberately sidesteps the rules of representative sampling and instead defines reproducibility as a local test of precision.

How, for example, do we respond to someone who asks the obvious, but embarrassing question? Is it possible to demonstrate reproducibility without considering accuracy? Although the correct answer is no, the standard definition says yes.

This is a typical example of the mischief created by reductionist theory. It strains our credibility by forcing us to assume indefensible positions. The new definition, which deliberately avoids the uncertainties implicit in the standard definition, copies biology by defining reproducibility as the same event that occurs locally and globally.

Let's see what happens when apply the two definitions of reproducibility to the amygdala as it exists in the IBVD. If we assume that all the points graphed in Figure 4.15 are correct, then it seems fair to conclude that the amygdala displays a considerable degree of biological variation.

Biology, however, requires both reproducibility (precision and accuracy) and adaptability (biological variation). But how does biology install - reproducibly – an accurate version of the amygdala - in each human brain, without having to restrict its size?

It defines reproducibility in the amygdala as the ratio of two volumes – the left amygdala (0.4 cm^3) versus the right (0.5 cm^3). If we start with the original 58 data points shown in Figure 4.15, plot the volume of left side against that of the right, we get the result shown in Figure 4.16. The disorderly cloud of points condenses onto a single point (expressed here as a decimal repertoire value). By aligning our *New Definition* to biology's left-right reproducibility rule for the amygdala (0.4:0.5), we discover that the same measure of reproducibility, precision, and accuracy apply to both local and global levels. By emulating biology, biology rewards us by giving us access to its highly prized package of skills.

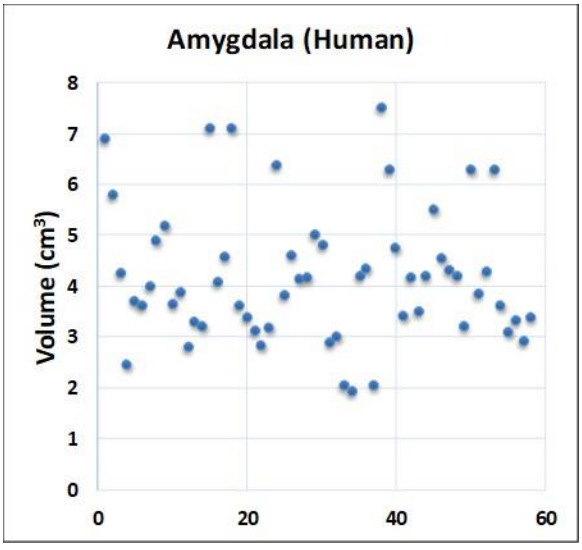
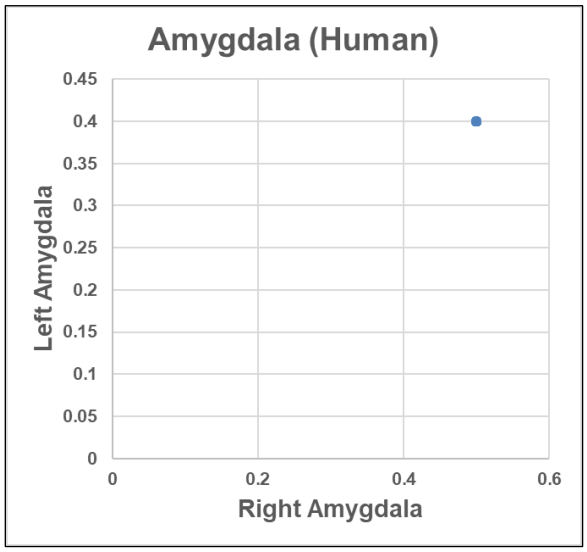


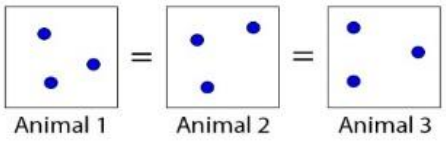
Figure 4.15 A plot of 58 estimates for the volume of the amygdala produces a cloud of data points (IBVD). Note that each data point represents the average of several patients (Original data adapted from Kennedy et al., 2012; From Bolender, 2012).



cit_n	mathematical marker
42	amygdala1amygdalaleft0.4amygdalaright0.5
50	amygdala1amygdalaleft0.4amygdalaright0.5
65	amygdala1amygdalaleft0.4amygdalaright0.5
78	amygdala1amygdalaleft0.4amygdalaright0.5
126	amygdala1amygdalaleft0.4amygdalaright0.5
137	amygdala1amygdalaleft0.4amygdalaright0.5
178	amygdala1amygdalaleft0.4amygdalaright0.5
234	amygdala1amygdalaleft0.4amygdalaright0.5
308	amygdala1amygdalaleft0.4amygdalaright0.5
314	amygdala1amygdalaleft0.4amygdalaright0.5
316	amygdala1amygdalaleft0.4amygdalaright0.5
329	amygdala1amygdalaleft0.4amygdalaright0.5
441	amygdala1amygdalaleft0.4amygdalaright0.5
453	amygdala1amygdalaleft0.4amygdalaright0.5
456	amygdala1amygdalaleft0.4amygdalaright0.5

Figure 4.16 When the left and right sides of the amygdala are expressed as 58 ratios, the scatter plot shown in Figure 24 is replaced by a single point (a decimal ratio value); (Original data adapted from Kennedy et al., 2012; From Bolender, 2012). The Excel table displays 15 of the 58 rows of data that demonstrate global reproducibility using level 1 complexity. Note that all the markers in the list come from different citations (cit_nu)

4.4 Level 2 – Biochemical Homogeneity



The postulate of biochemical homogeneity assures us that the members of a given population (e.g., organelle membranes) have the same biochemical composition, whereas the postulate of single location states that each constituent [i.e., marker enzyme] is restricted to a single cellular site (e.g., the endoplasmic reticulum). To wit, a marker enzyme (biochemical constituent) distributes uniformly at a unique cellular location (morphological component).

Recall that the postulate of biochemical homogeneity is captured theoretically and empirically as the Structure-Function Rule: $f(x) = mx$.

4.4.1 Structure-Function Rule and Biochemical Homogeneity

To test the postulates of biochemical homogeneity, we can plot the surface area of the endoplasmic reticulum (ER) of hepatocytes against the activities of several ER bound marker enzymes (Figure 4.17; Figures 4.18 to 4.21), as described earlier (Bolender, 2017). Notice that these relationships of structure to function define ratios, which we can detect with linear equations that display $R^2 = 1$ or ≈ 1 and pass through the origin (Figures 4.18 to 4.21). These equations predict membrane surface areas from enzyme activities of individual animals and adhere to the postulates of biochemical homogeneity. [Note: the

membrane surface areas were corrected for the section thickness biases according to Weibel and Paumgartner (1978).]

What can we learn about reproducibility from Figures 4.18 to 4.21? The spread of the three blue points along the regression lines in these figures identify the biological variation, but each point, which represents an individual animal, defines the same ratio (the slope: $\Delta y/\Delta x$) and generates the same equation (e.g., Figure 4.18: $y = 0.1677x$). This shows how the relationship of structure to function defines a **highly reproducible** property of biological systems. To biology, reproducibility exists as families of rules, which exist in our parallel complexities as equations. (e.g., Figures 4.18 to 4.21).

G-6-PASE		membrane	enzyme
ER		surface	activity
		m^2/g liver	U/g liver
		tissue	**homogenate**
		er	**g-6-pase**
animal 1	er-1	**4.870**	**29.031**
animal 2	er-2	**4.310**	**25.692**
animal 3	er-3	**4.620**	**27.540**

Figure 4.17 Data conforming to the postulate of biochemical homogeneity. This required allocating – proportionately – the average enzyme activity to the membrane surface areas of individual animals (Original data adapted from Bolender, et al., 1978 (Paper 1); From Bolender, 2017).

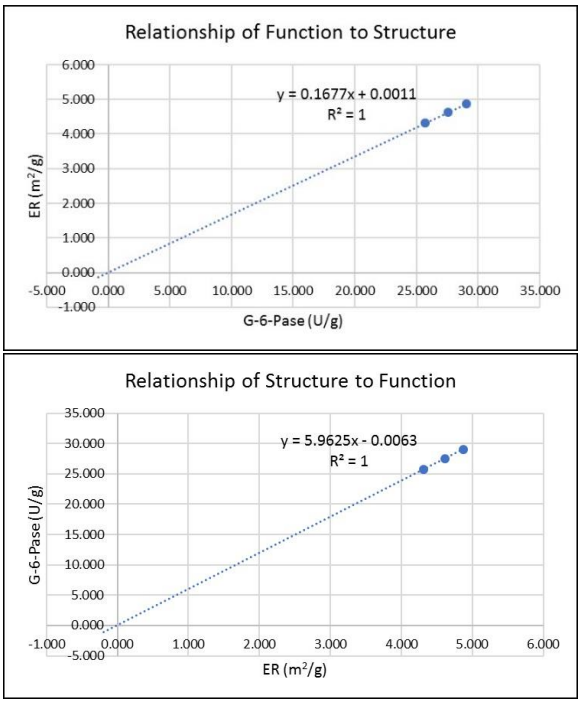


Figure 4.18 The $R^2 = 1$ equations identify the relationship of the surface area of the ER to the biochemical activity of G-6-Pase assayed in tissue homogenates (E+N). (Original data adapted from Bolender, et al., 1978 (Paper 1); From Bolender, 2017).

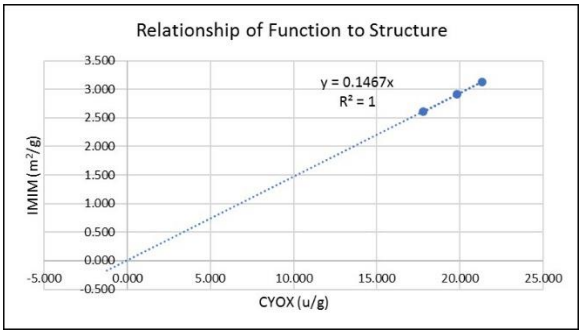


Figure 4.19 The equation identifies the relationship of the surface area of the inner mitochondrial membrane (IMIM) to cytochrome oxidase (CYOX) activity assayed in tissue

homogenates (E+N); (Original data adapted from Bolender, et al., 1978 (Paper 1); From Bolender, 2017).

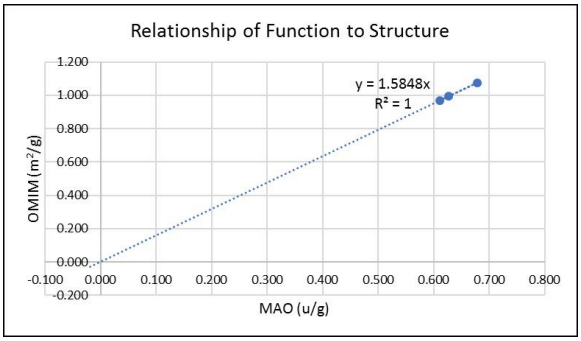


Figure 4.20 The equation identifies the relationship of the surface area of the outer mitochondrial membrane (OMIM) to monoamine oxidase (MAO) activity assayed in tissue homogenates (E+N); (Original data adapted from Bolender, et al., 1987; Bolender, 2017).

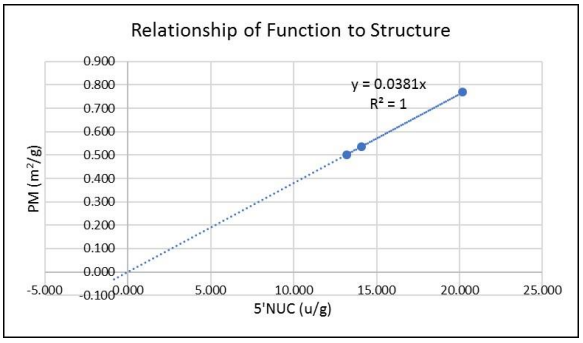


Figure 4.21 The equation identifies the relationship of the surface area of the plasma membrane (PM) to 5'nucleotidase (5'NUC) activity assayed in tissue homogenates (E+N); (Original data adapted from Bolender, et al., 1978 (Paper 1)*; From Bolender, 2017).*

What do these four figures tell us? We need biological variation to discover biology's ***structure-function rule***, which, in turn, defines reproducibility as individual data points sitting on the same regression line with $R^2 = 1$ or ≈ 1.

4.4.2 Reproducing Equations

Figures 4.18 to 4.21 translate the postulate of biochemical homogeneity into equations. But how do we verify these equations?

If these equations capture biological rules, then a similar data set in another publication should capture the same rules (equations). Let's test this supposition.

When we know an ER marker enzyme activity and its structure-function equation (Figure 2.4), for example, we can use this information to predict the surface area of the ER membranes from enzyme activities in each of several tissue fractions published in a different paper. The results appear in Figure 4.22. Notice that the equation has an $R^2 = 1$ and passes through the origin as predicted by the structure-function rule (Equation 2.4).

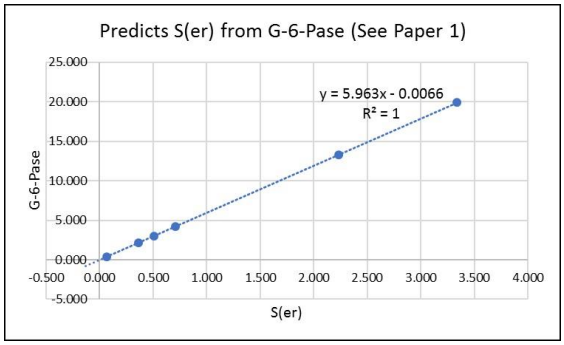


Figure 4.22 The equation of Paper 1 (Bolender et al., 1978; Figure 2.4) was used to predict ER surface areas from the biochemical data (G-6-Pase) of the tissue homogenate (H) and tissue fractions (E, N, M, L, P) in Paper 3 (Bolender et al., 1980). The result is a curve with an $R^2 = 1$ and effectively passing through the origin. Recall that de Duve's postulates and Equation 2.4 predict such a result. From Bolender, 2017.

This gives us two equations capturing the same rule, one from Figure 2.4 (paper 1) and the other from predicted ER surface areas (paper 3):

Original (Paper 1): $y = 5.9630x - 0.0066$ (4.1)
Predicted (Paper 3): $y = 5.9625x - 0.0066$ (4.2)

Since the equation of Paper 3 duplicates that of Paper 1, we can see that the same rule applies to both studies. Such a result suggests that capturing biological rules with equations represents a reliable measure of precision, prediction, reproducibility, and to a lesser extent accuracy (recall that stereological

estimates typically carry volume distortions). Moreover, the successful reproducibility test shown in Figure 4.22 offers further empirical proof in support of the postulates of biochemical homogeneity.

4.4.3 Reproducing Experimental Results

Next, we can try to reproduce the results of an experiment using the standard approach based on parts data and the new one based on copying and applying biology's rules. To this end, we will compare results published in two different papers (Paper 1: Bolender, et al., 1978; Paper 2: Losa et al., 1978).

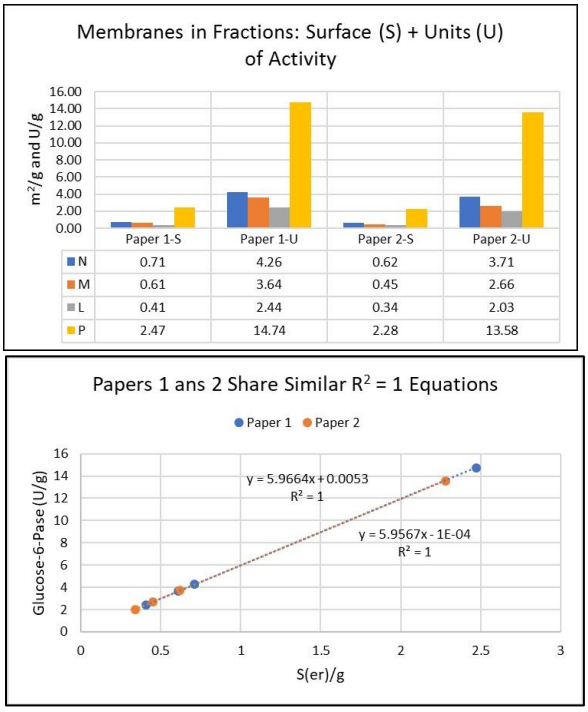


Figure 4.23 The plots Illustrate that the same mathematical relationship exists between the data of Papers 1 and 2 (ER surface vs. activity of its marker enzyme G-6-Pase) because the two curves overlap. Original data adapted from Bolender et al., 1978 (Paper 1), Losa et al., 1978 (Paper 2); From Bolender, 2017.

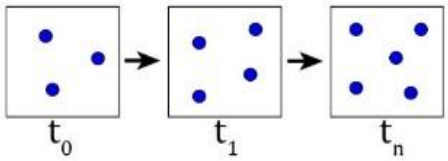
As shown in the top panel of Figure 4.23, the biochemical and morphological data – expressed as isolated parts - carry the expected biological variation, with comparable estimates differing by as much as 36%. If, however, we express each result as a structure-function equation, the two papers generate remarkably similar equations and overlapping curves:

Paper 1: $y = 5.9664x + .0053$ (4.3)
Paper 2: $y = 5.9567x - 1E - 04$ (4.4)

This tells us that the results of Paper 2 reproduce the results of Paper 1.

Why is this example useful? We're building a case to support the argument that rule-based equations are reproducible. We accomplished this by showing that the relationship of structure to function – defined by an organelle and a constituent marker enzyme – remained constant across different sets of animals repeatedly. By defining and demonstrating reproducibility with equations operating locally and globally (Papers, 1, 2, and 3), we can predict outcomes with acceptable levels of confidence (Equations 4.1 to 4.4). Since prediction is an expectation of a quantitative science, copying biology's package of skills is proving to be a promising strategy.

4.5 Level 3 – Organelle Changes



A change in a biochemical homogeneity can be expressed as an enzyme density (ED) wherein units of enzyme activity are related to 1 m^2 of ER surface area. [Consider, for example, the three squares shown above ($t_0 \rightarrow t_1 \rightarrow t_n$); the blue dots indicate that the ED increases from 3 to 5 Units/Surface.]

4.5.1 Enzyme Density Changes

Figure 4.24 shows how the enzyme densities of three ER marker enzymes of hepatocytes change in response to treatment with phenobarbital. Notice that the drug induced two of the three enzymes, but at different rates and that the enzyme densities fit linear curves with R^2s equal to or close to one.

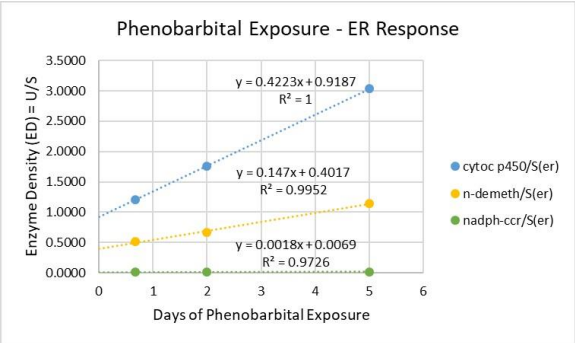


Figure 4.24 By expressing changes in membrane surface areas and enzyme activities with enzyme densities, distinct relationships of structure to function become apparent. Notice that the all three enzymes use the same general change rule (f(x) = mx + b), wherein the structure-function change equation serves as a higher-level rule for the variable enzyme densities (the dots). In effect, two rules combine to change the relationship of structure to function in the ER membranes (Original data adapted from Stäubli et al., 1969; Bolender, 2018).

At level 3 complexity where we detect organelle changes, our measure of reproducibility continues to derive from biology's rules that translate into equations with R^2s = 1 or ≈ 1. Accordingly, rule-based equations not individual data points become the preferred measure of reproducibility.

4.5.2 Reproducibility as a Nested Complexity

For biology, reproducibility involves far more than simply duplicating an experimental result. Instead, it would validate a change by demonstrating that each of several individuals respond to a given change by applying the same set of rules. Let's look at an example.

If we select the enzyme density curve for cytochrome P450 from Figure 4.24, we can run a reproducibility test on the ER membranes in hepatocytes as they change.

In Figure 4.25, each ED data point defining the cytochrome P450 curve represents the average value for 5 animals. If, instead, we plotted the individual animal data, we would expect all 15 data points to produce the same curve (analogous to what we found in Figures 4.18 to 4.21).

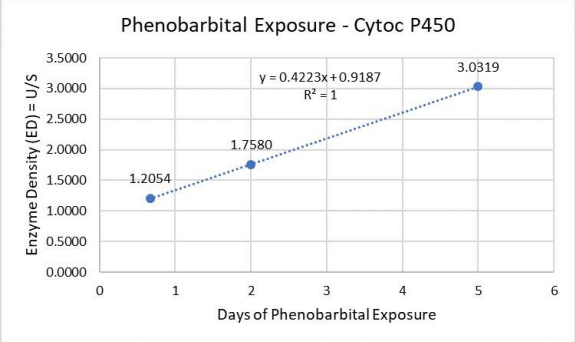


Figure 4.25 The curve represents the hepatocytic response to phenobarbital (100 mg/kg body weight/day) over 5 days of exposure (Original data adapted from Stäubli et al., 1969).

The reproducibility test of interest to the hepatocytes – in each animal – was to produce enzyme densities – over time - that obeyed a structure-function change rule consistent with surviving the consequences of an exposure to a xenobiotic (phenobarbital). Think of it this way. Of the 15 animals given the same problem to solve, all 15 came up with the same equation (4.5) obeying the same rule:

Structure-Function Change Rule

$$\boldsymbol{f(x) = mx + b};$$

$$y = (0.4223x) + 0.9187. \qquad (4.5)$$

$$ED_{PB(day\ 5)} = (0.4223 \times 5\ days) + 0.9187.$$

$$ED_{PB\ (day\ 5)} = 3.03\ U/m^2$$

This tells us that the response of hepatocytes to phenobarbital is a rule-based event and as such reproducible.

However, the structure-function change rule of Figure 4.25 is the result of an underlying set of structure-function rules (Equations 4.6 to 4.9); as shown in Figure 4.26.

Structure-Function Rule

$$\boldsymbol{y = mx};\ \boldsymbol{m = \frac{y}{x}};\ \boldsymbol{ED = \frac{y}{x}}$$

$$ED_{PB(day\ 5)} = \frac{U}{S} = 3.0319\ U/m^2 \qquad (4.6)$$

$$ED_{PB(day\ 2)} = \frac{U}{S} = 1.7580\ U/m^2 \qquad (4.7)$$

$$ED_{PB(day\ 0.67)} = \frac{U}{S} = 1.2054\ U/m^2 \qquad (4.8)$$

$$ED_{Control(day\ 0)} = \frac{U}{S} = 1.0731\ U/m^2 \qquad (4.9)$$

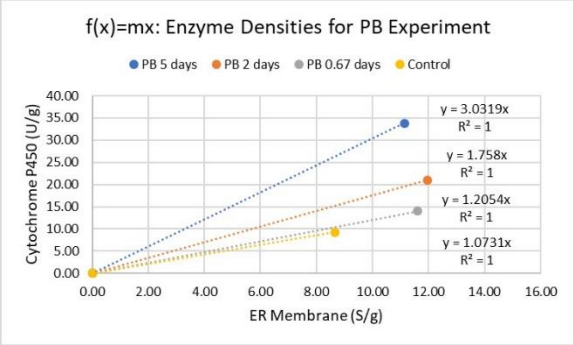


Figure 4.26 The rules responsible for changes in the ER membranes map back to the structure-function rules (Original data adapted from Stäubli et al., 1969).

In Figure 4.26, the biological test of reproducibility requires that all 5 animals contributing to each data point (days 0 to 5) to produce ER membranes having the same concentration of cytochrome P450 (EDs) at roughly the same time (recall Figures 4.18 to 4.21).

Figures 4.25 and 4.26 show us how a cell relies on reproducibility to build one level of complexity onto another. We can see this in the way the structure-function change rule depends on the success of the structure-function rule. In effect, reproducibility makes biological complexity possible and vice versa.

Let's add up the score. The results shown in Figure 4.25 include a wealth of reproducibility events; 15 animals duplicated the change equation, whereas 5 animals duplicated the structure-function equation at each of 3 days in Figure 4.26. This gives us a total of 30 examples of reproducibility for cytochrome P450. Include the other two enzymes (n-demethylase and NADPH cytochrome c reductase), and the total reproducibility score for Figure 4.24 comes to 90. If we include the control, the score jumps to 95. Basically, reproducibility exists as a ubiquitous property of biology. It quickly disappears, however, when we mix pure biological data with methodological variations and throw away the rule-based connections between structure and function.

4.5.3 Reproducibility as a Clue

Consider the 15 animals (5 per time point) used to generate Figure 4.24. For each time point for every enzyme, the hepatocytes of each animal came up with more or less the same structure-function and structure-function change equations in response to the phenobarbital challenge. Moreover, all the animals solved the problem during the first day of exposure and anticipated the data points and slopes of all three curves. How else could they have generated the equations with R^2s = 1 or ≈ 1 equations?

But, how did they manage such mathematical magic? Is this a biological phenomenon wherein groups of animals (and their hepatocytes) share spatial or temporal proximity such that the state of a given cell (or animal) cannot be determined independently – but only for the biological system as a whole? Are cells and animals - like their underlying

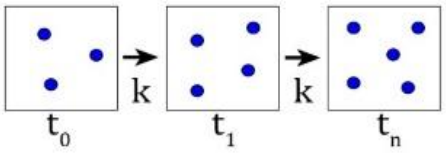
particles - entangled? Alternatively, hepatocytes might just be very good at solving difficult optimization problems by applying the same rules under a given set of circumstances.

In the first scenario, we might test for quantum entanglement at the level of cells, whereas the second one would require catching a cell in the process of solving a very difficult mathematical problem. The second option would seem to be more consistent with expectations of a primer.

Recently, it was reported that an amoeba can solve a travelling salesman problem - involving eight cities - faster than we can (Zhu et al., 2018). Since most would agree that such a feat is mathematically remarkable, we can infer that cells such as hepatocytes have what it takes to figure out and reproduce the results shown in Figure 4.24. If, in fact, cells are mathematical wizards, then one day we might discover that each enzyme density shown in the figure represents an optimal solution to a linear programming problem. However, since one equation ($f(x) = mx$) is embedded in another ($f(x) = mx + b$), the hepatocytes might be solving multiple optimization problems simultaneously.

What's biology trying to tell us? Should we be focusing our attention and resources on detecting and reproducing black box changes to assuage the demand for better P values or should we be copying and profiting from biology's approaches to advanced mathematics? Is biology needling us by offering a Hobson's Choice? To wit: "Either pursue biology as a complexity or use complexity to pursue biology."

4.6 Level 4 – Rates of Change



By fitting enzyme densities to a linear curve (zeroth order), the slope of the curve (m) becomes the rate constant k - the rate of change.

4.6.1 Rate Constants

Rate constant equations allow us to predict missing data points and hunt for similar equations. Moreover, reproducibility and prediction become interchangeable when we know the rules biology is using to bring about a change.

Figure 4.27 uses zeroth order rate equations ($[A_i] = [A_0] + kt$) to calculate the missing time points seen in Figure 4.24 at days 3 and 4. The slopes of the curves, which identify unique rates of enzyme synthesis, exist under tight controls wherein the R^2s are equal to or close to 1.

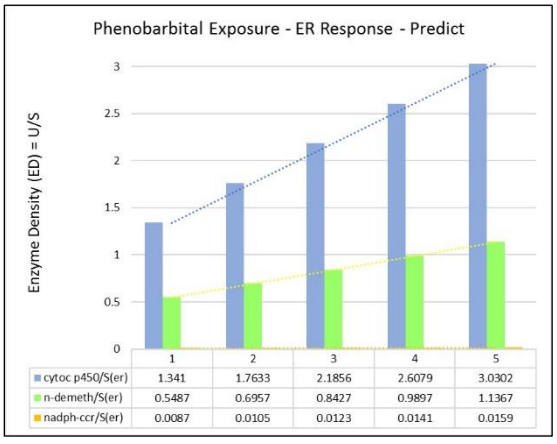
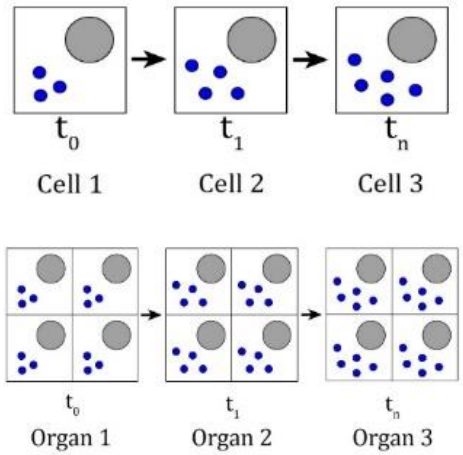


Figure 4.27 Using enzyme densities and rate constant equations, the missing data points for days 3 and 4 were predicted (Original data adapted from Stäubli et al., 1969; Bolender, 2018).

By knowing the rate constants for multiple cytoplasmic organelles and the relationships of one constant to another, we can begin to unravel mathematically the complexity surrounding the phenotypic responses of cells.

4.7 Levels 5, 6 – Cell and Organ Changes



The reference we choose for our data has a direct bearing on how we set our reproducibility arguments. When expressed as an enzyme density, we now know that we can interpret a biochemical homogeneity at multiple levels of complexity: organ, cell, organelle, and molecule.

4.7.1 Biased Results

One of the advantages of interpreting biological data with equations is that we can see how our choice of data reference influences the results. If we assay a membrane bound marker enzyme (cytochrome p450) and relate it to the total liver weight and to a gram of liver, we can get results, for example, that differ by as much as 70% (Figure 4.28). Why? In biology, data references that we routinely assume to be constants often choose to behave as variables. Consequently, they often generate faulty and misleading results.

Consider the following example of a typical concentration trap. When exposed to phenobarbital, hepatocytes synthesize substantial amounts of new ER membranes containing drug-metabolizing enzymes. Consequently, the hepatocytes get larger and fewer of them can fit into a gram of liver. When compared to controls, the experimental time points report enzyme activity coming from fewer and fewer hepatocytes as the exposure continues over time (Figure 4.28). Recall that such data become ambiguous because the data for each time point comes from a different number of hepatocytes. As a result, experimental outcomes based on biological concentrations are likely to be correct only about 50% of the time – even when differences are reported to be statistically significant (see Chapter 6).

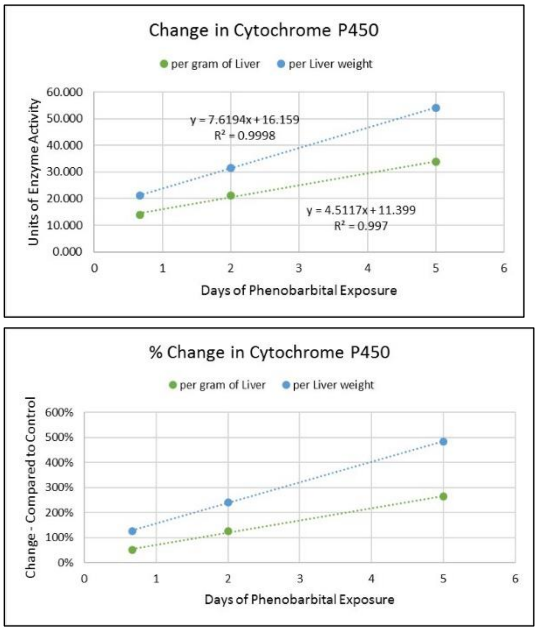


Figure 4.28 Collecting data from the same number of hepatocytes (per liver) detected almost twice as much change - compared to the widely used gram of liver reference (Original data adapted from Stäubli et al., 1969; Bolender, 2018).

4.7.2 Unbiased Results

We can avoid a heavily biased result by relating concentration data (e.g., enzyme activity/g liver) to an average cell or to the weight of the liver. Recall that changes expressed per average cell or per organ produce the same relative results (detected as a similar amount of change).

Since counting hepatocytes with stereology was not a viable option for the experimental studies under consideration (Bolender, 2018), absolute data derived from the organ weight became the method of choice. For the enzyme data, this calculation consists of multiplying the liver weight (g) by the enzyme concentration - the units of enzyme activity per gram (U/g):

$$U_{liver} = U/g\ liver \times W_{liver\ (g)}. \qquad (4.10)$$

Although this approach works reasonably well for biochemical data, it becomes suspect when applied to the membrane surface area estimates of stereology. Why? Because a cm^3 of living tissue and a cm^3 of fixed and embedded tissue do not necessarily contain the same number of cells. Recall that stereological concentrations often carry biases related to volume distortions (Figure 3.17 and Chapter 5).

A correction equation (CCC) mitigates both the distortion problem of stereology and the cell number problem of both methods by keeping the original number of cells in a cm^3 of liver constant. This is accomplished by allowing the cm^3 of liver to "enlarge" in proportion to the enlargement of the hepatocytes, which are assumed to be wholly responsible for the changes in liver volume (or weight).

After relating both the morphological and biochemical data to the total liver weight, the recalculated enzyme densities (Figure 4.29) were found to agree with the original estimates (Figure 4.24), which were calculated using a gram of liver. In this case, equations were used to check on the reproducibility of a result after the corrections (CCC) were applied.

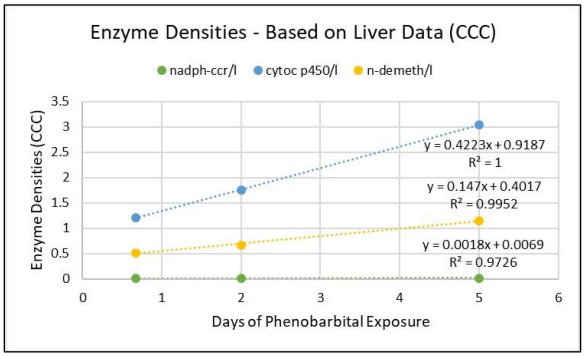


Figure 4.29 The results shown in Figures 4.24 and 4.27 are the same (Original data adapted from Stäubli et al., 1969; Bolender, 2018).

In this chapter, the focus of the reproducibility test shifted from duplicating data points to duplicating first principles, rules, and equations. This was accomplished by copying biology's approach to reproducibility, wherein the same result is expected at both the local and global levels.

4.8 Solving Reproducibility

Solving reproducibility begins by understanding what it brings to the sciences. If we agree that it includes an ability to repeat the results of an experiment, then we are on the same page. But what does a test of reproducibility test? Physicists and chemists, who view their science as a body of knowledge derived from first principles, theorize the existence of a principle, express it as an equation, and then test the validity of their idea empirically by running experiments. Here the primary focus of our reproducibility test is not just to duplicate the experimental result, but to confirm the existence of underlying principles.

Currently, the biological sciences have reinterpreted the reproducibility test of the physical sciences to include just the final experimental result. In effect, we have become

willing to accept what comes out of a “black box” without the usual concern for what’s going on inside the box (the underlying principles). Consequently, when challenged, the best we can do is bemoan the presence of a reproducibility crisis (Baker, 2016).

But we still need to ask the question basic to our understanding of reproducibility. To wit, why does reproducibility work for the physical sciences within a theory structure based on reductionism, but this is not the case for biology? The physical sciences succeed wonderfully within the framework of reductionism because they play in a one table league, which is based on the elements of the periodic table.

Biology, on the other hand, suffers grievously under the simplifications of reductionism because it must play in a two-table league. In addition to the periodic table, biology must also deal with its enigmatic gene table, which adds adaptability, allows rules to change, and involves nested complexities.

This explains why our definition of reproducibility must be held to the higher standard of physics and chemistry. To wit, if reproducibility cannot be demonstrated both locally and globally according to rule, it doesn’t count.

CHAPTER 5

DATA

The single, most difficult part of solving biology requires going to the literature, extracting data, and reformatting them in ways that provide access to biology's rules and first principles. Rules and principles, which are highly reproducible, transform an otherwise chaotic literature into a well-oiled machine capable of diagnosis, prediction, and complex problem solving.

On one thing we can all agree. Our research efforts quickly become counterproductive when they fail to deliver results with acceptable levels of confidence. Since evidence for these shortfalls continues to accumulate at an alarming pace, it seems reasonable to suggest that we may need to rethink many of our experimental approaches. Such a task begins by asking the telling question. If biology consistently delivers its results with precision, accuracy, and reproducibility, why can't we do the same? What's stopping us? At this point in our story, absolutely nothing.

Getting help from biology requires a two-step approach. After taking biology apart, we must put it back together. The primer argues that this includes little more than giving biology back its connections, rules, and complexity. But, why?

Biology runs it business as a complexity with rules defined mathematically, the evidence for which appears to exist largely in its connections and relationships of structure to function. By throwing away its connectivity, however, we unintentionally prevent biology from telling us what we need to know. Because of this questionable practice, we as community must now deal with the ongoing crisis of confidence in our ability to do our job responsibly.

To get our research data to qualify as precise, accurate, and reproductible, we need to identify new experimental strategies and data types capable of delivering such results. In this chapter, we summarize some of the options now open to us.

5.1 UNBIASED SAMPLING

Unbiased sampling methods guarantee that every part in play has an equal chance of being sampled. This applies to parts of all sizes, extending from microscopic to macroscopic from molecules to organisms. A sample becomes representative ***only*** when it faithfully replicates all the parts of interest as they exist in the parent structure.

The point of unbiased sampling is to obtain a small, but representative sample that can be extrapolated back to the original material. Such a procedure represents the first step toward providing dependable and reproducible estimates. The method of sampling, however, determines what types of information can be recovered. Tissue homogenization provides an unbiased sample, but the process forfeits most of the structural information and we are left largely with total averages. In contrast, sectioning intact tissues forfeits dimensional information that must be recovered with serial section reconstructions or stereological (probabilistic) methods. In both cases, an adequate recovery depends on the validity of experimental designs, methodological biases, data types, and methods of analysis.

Let's look at a few issues related to homogenization and tissue sectioning.

5.1.1 Homogenizing Cells and Tissues (Structural Order Minimized)

The structural integrity of the original object containing the parts of interest is lost by homogenizing cells and tissues. This represents the standard reductionist approach to studying molecules and organelles, wherein biology is reduced to parts, which can be isolated, concentrated, and characterized.

An intact tissue or a collection of cells is homogenized, and an aliquot taken for analysis. Such an aliquot represents an unbiased sample, provided the estimate can be extrapolated back to the original material. In practice, the assay is repeated several times and the average taken. If the homogenate is fractionated, then the rules of analytical fractionation apply, wherein both recoveries and balance sheets will be needed to extrapolate and validate the data (de Duve, 1974). Isolating parts and relating them to a mg of protein, however, may not always satisfy the unbiased sampling requirement. [Variations in decanting supernatants, for example, influence the amount of protein in a tissue fraction, which affects estimates for enzyme activities related to a mg of protein.]

5.1.1.1 Advantages

- **Access to molecules:** Homogenizing biological material provides access to molecules otherwise inaccessible in structural compartments.
- **Convenient counting of molecules:** By extracting the molecules and suspending them in solution, their concentration can be determined using optical density (OD) methods (e.g., absorbance, transmission, etc.).
- **Automation:** Collecting and analyzing the large amounts of data created by experiments in genetics, molecular biology, and biochemistry have been well-served by automation.
- **Formation of data ratios:** One of the best sources of order in biology exists in relationship of one part to another. When two parts (e.g., molecules) are related to the same data reference (a volume, surface, mg protein, gram of tissue, et cetera), the reference variable cancels out, leaving a dimensionless ratio. Forming such ratios minimizes methodological biases and animal variability – provided that each step of the experimental process is unbiased.

5.1.1.2 Caveats

- **Forfeits structural information:** Since homogenization can disrupt the identifiable structural locations of a molecule, our ability to reconstruct the activities of that molecule as it exists in the original tissue can become problematic. Recall that changes in a molecule of interest – located at one or more structural locations - will be detected as a single, total value referenced to its parent structure. If, however, the total number of cells of interest changes, then the interpretation of the results can become questionable. A change can be the result of a change in the cells, a change in the number of cells contributing data, or some combination of the two events.
- **Restricts biological engineering:** Engineering biology (reverse and forward) depends importantly on structural information distributed throughout the biological hierarchy of size. A homogenate no longer contains this original information. In short, reconstructing biological complexities requires both biochemistry and intact tissue morphology.
- **Introduces ambiguity:** An optical density represents a concentration, namely, the

number (N) of molecules contained within a unit of reference volume (V) or surface area (S):

$$OD = N/V \ \ or \ OD = N/A. \qquad (5.1)$$

In biology, an optical density detects an event that can be influenced by changes occurring both in the numerator and denominator of the concentration ratio and by the design of the experiment. Recall that the data reference (the contents of the containing volume) is expected to be different for each experimental setting. Therefore, optical densities – as isolated units of information – can quickly become ambiguous and misleading.

One of the easiest ways of getting into trouble quickly is to rely exclusively on concentration data (i.e., optical densities) for detecting biological changes (Figure 7.25). Such a practice requires two risky assumptions: the average cell size and the total number of cells of interest must remain constant, otherwise the data become ambiguous. Dividing experimental by control concentrations becomes a high-risk operation because the data references are likely to be incompatible (i.e., the data reference fails to cancel out). See chapter 7 for a worked example.

- **Data Inconsistencies:** Biochemical assays and results tend to vary across laboratories (Figure 7.44). Moreover, data related to a mg of protein adds to the confusion because it too varies across laboratories (Bolender, 2017).

5.1.2. Sectioning or Slicing Intact Cells and Tissues (Structural Order Maximized)

The structural organization and representativeness of the original object is largely retained when designed-based sampling methods and tissue sectioning generate unbiased samples analogous to those coming from the tissue homogenates of biochemistry.

5.1.2.1 Advantages

- **Provides ready access to parts and connections:** Morphology supports a mathematical biology in that all the visible parts can be quantified and connected throughout the biological hierarchy of size.
- **Supports engineering (reverse and forward):** Data collected from sections can contribute significantly to reverse and forwarding engineering.

5.1.2.2 Caveats

- **Generates Biases:** Tissue preparation, sectioning, staining, section thickness, and data collection (point and intersection counting) can contribute biases - detectable or undetectable. Strategies for minimizing the effects of such biases exist and suggested workarounds can be found in the EBSP Reports (Bolender, 2001-2006; 2013).

5.2 The Data Problem

When we take biology apart, we end up with a collection of parts that can be characterized in terms of their composition and properties. To understand what's happening to these parts when they change in a biological setting, however, the interrelationships of the parts

must be reestablished in accord with biology's rules.

Our purpose here is to consider two types of data: simple (the result of reducing biology to a collection of parts) and complex (the result of reconstituting biology from simple data).

5.3 Simple Data Types

Simple data types include concentrations and absolute values. Such data, which involve parts but not connections, constitute only a small fraction of biology's information inventory. Consequently, parts data – by themselves – are severely limited in what they can tell us about the rules and principles under which biology operates. It follows that limited data limit progress.

Because simple data are compelled by our methods to carry a considerable load of biological variation and experimental bias, results derived therefrom tend to produce outcomes with low levels of confidence (Figure 5.1). The typical responses to such shortcomings include increasing sample sizes to detect changes, introducing unbiased sampling methods, correcting for methodological biases, and replacing concentrations with absolute values. As Figure 5.1 suggests, simple data carry low levels of confidence because they are largely incapable of gaining access to biology's rules and first principles. Confidence comes from biology first, then from statistics.

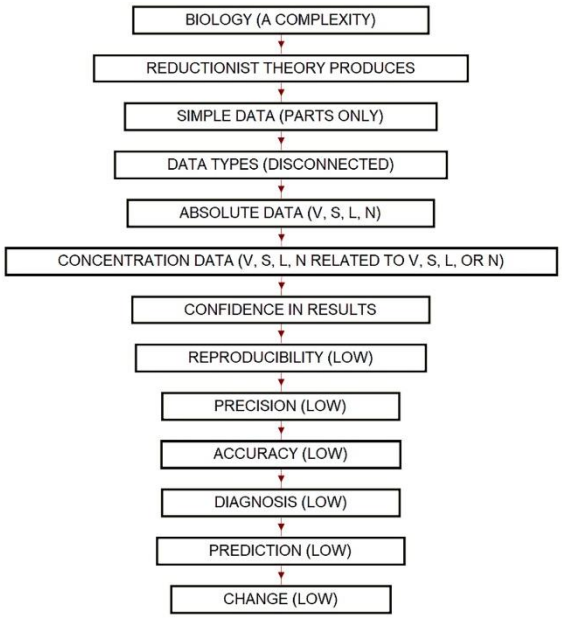


Figure 5.1 The reductionist approach to experimental biology reduces a complexity to a simplicity by taking it apart. The resulting parts, which we can characterize quantitively as volumes (V), surface areas (S), lengths (L), and numbers (N), can be reported as absolute values or concentrations. Missing from this scheme are the connections and the complexities – two of the three main factors that define biology.

5.3.1 Include Concentration and Absolute Data

Since we typically collect data as concentrations, which can have unstable reference volumes (Figures 6.3, 6.6, 7.2, 7.24, 7.25), detecting biological changes often requires the additional step of converting concentrations into absolute values. Problems arise, however, in that the methods used to make this conversion often rest on shaky assumptions. Why?

In biochemistry, for example, we can calculate an absolute value for an enzyme activity by multiplying a concentration (units of activity/gram of tissue) times the total weight of the tissue:

$$Absolute\ Value_{enz} = Weight_{tissue} \times \frac{Activity}{g\ tissue} \quad (5.2)$$

In an experimental setting, wherein we look for a change in an activity, multiple factors can impact the change. The activity in the cells can change, as well as the numbers, shapes, and sizes of cells – all of which can influence the result. Since Equation 5.2 uses only two pieces of information, the tissue weight and the activity per gram of tissue, the absolute value delivers a black box result. It tells us what happened, but not how it happened. Furthermore, we learn nothing from such a result about the effects of the methodological biases.

An analogous situation exists for morphology with an added twist. We can estimate, for example, the surface area of the ER in liver hepatocytes using a hierarchy equation:

$$S_{ER} = V_{liver} \times \frac{V_{paren}}{V_{liver}} \times \frac{V_{hep}}{V_{paren}} \times \frac{S_{ER}}{V_{hep}} \quad , \quad (5.3)$$

but we are still left with the ambiguity associated with changes in cell numbers and sizes plus the confounding effects of specimen preparation.

In Equation 5.3, for example, we are multiplying the volume of fresh liver by three separate concentrations: two of which may come from light microscopy and a third from electron microscopy. This means that to evaluate equation 5.3, we must assume that the contents of a cm^3 of a part (e.g., $\mathbf{V_{hep}}$) in the numerator ($\mathbf{V_{hep}}/V_{paren}$) is the same as the corresponding one in the denominator ($S_{ER}/\mathbf{V_{hep}}$) – otherwise they will not cancel out. Since different preparations for light and electron microscopy produce different volume distortions, the same units in the numerator and denominator become incompatible and fail to cancel out. This means that we need to assume that we can evaluate Equation 5.3 (all denominators of the densities cancel) and that control and experimental data share similar biases (e.g., volume distortions and section thickness artifacts). Since such assumptions are unrealistic, they quickly become indefensible (recall Figures 3.16 to 3.18).

This inconsistency in reference volumes points to a fundamental flaw in the mathematical reasoning behind the use of hierarchy equations in biological stereology (Bolender, 2013). Alternatively, we can just concede that hierarchy equations such as 5.3 are likely to produce biased estimates. However, this diminishes the authority of our experimental results.

The best argument to make in defense of hierarchy equations is that we are applying the same preparative methods to both the control and experimental time points with the assumption that the volume distortions remain the same in both settings. Although such an argument may provide local cover for a given experiment, it works against our larger goal of reproducing experimental results globally.

The examples provided by Equations 5.2 and 5.3 help to explain why our experiments tend to work at a local level, but not globally. With multiple variables and biases in play, routinely demonstrating reproducibility within tight tolerances quickly becomes an unrealistic goal. Support for such a conclusion comes from a biology literature overflowing with inconsistencies and distressingly short on validity (Ioannidis, 2011).

5.3.2 Include Biased Data

In an experimental setting, absolute data (x) would appear to be more reliable than concentrations (x/y) for detecting biological changes because one variable is in play instead of two. However, control and experimental data are likely to carry multiple and often different methodological biases.

A worked example will help. Let's see how a single source of bias (section thickness) operates when we use stereology to estimate membrane surfaces from electron micrographs. Although stereological estimates for surface areas assume sections with no thickness, the sections used for electron microscopy (EM) are thick enough to introduce a wide range of overestimates (see Weibel and Paumgartner, 1978).

The absolute surface of the hepatocytic ER calculated with Equation 5.3, for example, overestimates the unbiased value by about 30% (Figure 5.2). Notice in the figure that membranes forming spheroidal shapes with diameters similar to that of the section thickness (SER) introduce the largest overestimates, whereas sheet-like arrangements of membranes account for the smallest (plasma membrane (PM)).

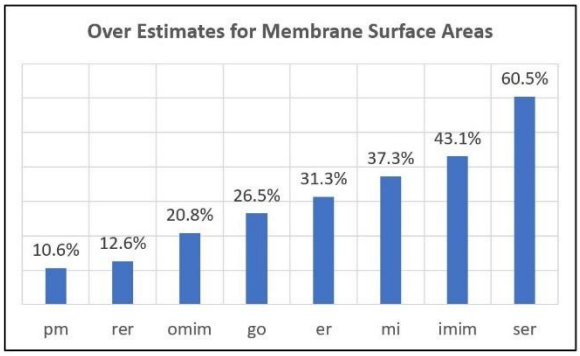


Figure 5.2 Surface areas determined from electron micrographs of hepatocytic membranes can overestimate the true value by as much as 60.5% (Original data adapted from Weibel and Paumgartner, 1987).

Now let's use the overestimates given in Figure 5.2 to see how the section thickness biases can impact the results of an experiment. We know, for example, that the ER shuttles membranes to all the compartments listed in the figure except for some constituents of the inner mitochondrial membrane (IMIM). [Recall that mitochondrial DNA codes for some of its membrane proteins.]

If we ignore the Weibel-Paumgartner corrections, we can get membranes to change their surface areas simply by moving them into a cell compartment where they assume a different size and or shape. For example, by transferring 1 m^2 of SER membrane to the plasma membrane (PM), the pm would increase its surface area by only 0.5 m^2 because the configuration of the membrane would shift from a high (SER) to a low (PM) section thickness bias. It gets worse. We would then have to explain what happened to the "missing" 0.5 m^2 of SER membrane surface area. The point? In the absence of section thickness corrections, biases remain unchecked and can produce all kinds of quantitative mischief.

Why is it important to know about such seemingly minute details? To get at the fundamentals of internal cell kinetics, we will need reliable biochemical densities and rate constants for the membranes and enzymes as they shuttle packets of information – encoded in membranes - throughout the cell. By minimizing the effects of biases, we maximize our chances of finding $R^2 = 1$ equations. In return for paying careful attention to the details, we can improve our access to both the kinetics and the rules needed to work out the underlying genetic and cellular mechanisms – upstream and down. Such information becomes critical when it becomes incumbent on us to explain the intended and unintended consequences of genetic engineering.

5.4 Complex Data Types

Figure 5.3 summarizes the complex data types, which rely on ratios, patterns, mathematical markers, connection ratios, relationships of structure to function, rules, and $R^2 = 1$ equations.

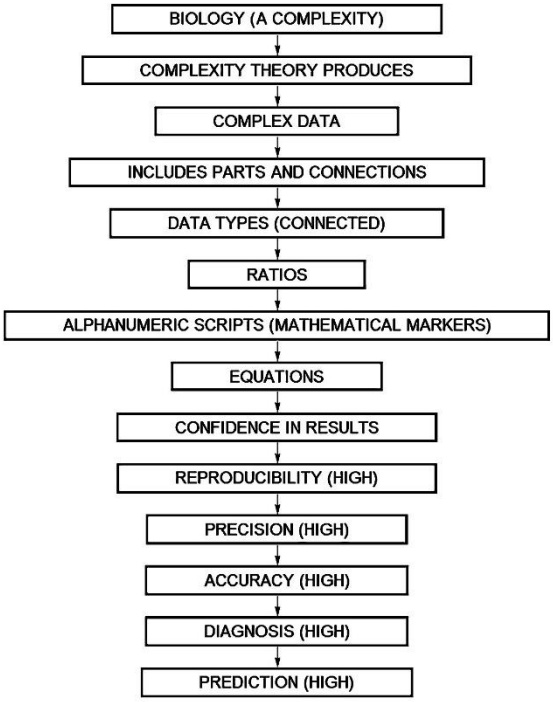


Figure 5.3 Complexity includes both parts and connections expressed as ratios. Data types, based on ratios, include mathematical markers, concentration ratios, and equations capable of delivering results with higher levels of confidence. Such data types contribute to the parallel complexities needed to copy biology.

Given the burden of assumptions, biological variation, and experimental biases, the likelihood of simple data types achieving widespread reproducibility at the levels some experts now consider acceptable (Colquhoun, 2014) seems remote. In a word, simple data dealing with simple questions is a construct far removed from the complex reality of biology. A failure to note this distinction is likely to result in the continued erosion of confidence and public support (see online: NIH Research Funding Trends).

5.4.1 Copy Biology

A primer bearing the title of "solving biology" has little choice but to come up with a more capable collection of data types, which address the concerns of the experts and put us on a path to finding durable solutions to many of the hard problems pressing us for solutions.

The strategy remains constant. Copy biology's game plan by coming up with data types running parallel to those of biology and then show that they derive from rules based on first principles.

The tricky part of copying biology involves shifting our frame of reference from simple to complex – both theoretically and empirically. This involves converting the simple data published in the literature into complex forms. These new data types include parts and connections, serve as universal data types, and identify rules as reproducible patterns that translate into $R^2 = 1$ and $R^2 \approx 1$ equations. Complex data types generate equations, equations reproducibility, and reproducibility engenders precision, accuracy, diagnosis, and prediction. A primary reason for embracing complexity is to demonstrate that we can deliver big solutions to big problems – by recruiting the expert help we need from biology.

5.4.2 Use Data Strings

Although biology controls whatever it needs to, it does so by allowing some rules to exhibit different properties when operating in local and global settings. Biology designs its local rules such that they allow the size of biological parts to vary from one organism to another with the purpose of accommodating individual differences, such as body size, nutrition, and external influences. Such differences, which we identify as biological variation, typically produce population averages lacking in the precision and accuracy needed to detect slight differences or to replicate experiments. Although our current training tends to focus on the properties of parts, biology sees things differently. It prefers

rules based on the ratios of its parts to define and maintain its precision, accuracy, and reproducibility. Biology succeeds because it makes use of everything in its package of skills (Figure I1).

Why? Coordinated gene expression, for example, includes the production of many different molecular types at the same time, with each molecule displaying its signature rate constant (e.g., Figure 4.24). At each time point, however, the relationship of one molecular type to another defines a ratio that continues to change until it reaches its new dynamic equilibrium (steady state). In effect, it changes its structural-functional state – analogous to a thermostat reaching and holding a specific temperature.

By replacing individual parts with strings of parts and connections, we can capture snippets of the rules being used by biology to turn its parts and connections into complexities. Such strings satisfy the requirements of a complex data type.

5.4.3 Convert Simple Data to Complex

At level 1 complexity, going from simple to complex begins at the level of individual publications. Parts involved in the same experiment and sharing the same data reference are rearranged into alphanumeric strings based on the ratios of the parts. Such strings (data pairs, triplets, quadruplets, et cetera) include mathematical markers and connection ratios as summarized below.

The letters a, b, c, d identify the names of parts and x, y, z, q the values of their corresponding ratios:

Data Pairs (2 parts, 1 connection, 1 ratio)

- ax:by
- a(x/x):b(y/x) = raw data
- a(x/x):b(y/x) = stored in decimal bins
- axby = mathematical marker (based on decimal values)
- partx: party = connection ratio
- partxparty = connection ratio

Data Triplets (3 parts, 2 connections, 2 ratios)

- ax:by:cz
- a(x/x):b(y/x):C(z/x) = raw data
- a(x/x):b(y/x):C(z/x) = stored in decimal bins
- axbycz = mathematical marker (based on decimal values)
- partx: party: partz = connection ratio
- partxpartypartz = connection ratio

Data Quadruplets (4 parts, 3 connections, 3 ratios)

- ax:by:cz:dq
- a(x/x):b(y/x):c(z/x):d(Q/x) = raw data
- a(x/x):b(y/x):c(z/x):d(Q/x) = stored in decimal bins
- axbyczdq = mathematical marker (based on decimal values)
- partx: party: partz: partq = connection ratio
- partxpartypartzpartq = connection ratio

This process of generating alphanumeric strings paper by paper produces universal biology databases wherein morphological and biochemical data share the same format. Since such data sets define distinct patterns, they can be quite effective at generating clues and solving problems related to change, diagnosis, prediction, and reproducibility.

5.4.4 Map Data

Data strings detect patterns and reproducibility quantitatively. Mathematical markers and connection ratios, which serve as major data types in universal biology databases (e.g., parallel complexities), extend the range of our frame of reference from local patterns to global.

By accommodating large data sets assembled from the literature, data mapping provides a broad overview of what's happening biologically. Moreover, it teaches us two important lessons. First, we learn that global data define reproducibility because they derive from local data. In effect, local and global rules are the same – one predicts the other. Second,

data strings can act as unique identifiers of specific phenotypic states in numbers large enough to display diagnostic and predictive properties. This tells us that the same objective approach to clinical diagnosis can be expected to work effectively for both individuals and populations.

5.4.5 Deliver Structure-Function Data

A biochemical density (DB), for example, relates units of biochemistry to a constant morphological reference (e.g., 1 m^2) by expressing it as a ratio of function to structure (F/S). For stereological estimates, the references least likely to be corrupted by volume distortions include surface areas and lengths. By leveraging mathematical connections existing between structure and function, biochemical densities allow us to access several levels of biological complexity – previously unavailable.

5.4.6 Generate Rate Constant Data

By defining the speed at which a concentration changes over time, a rate constant (K) allows us to predict data points (Equation 2.27) and to compare the behavior of several parts simultaneously as they exist in complex settings (Equation 2.18). By substituting enzyme densities for the typical biochemical concentrations, we can now calculate rate constants relative to an organelle *in situ*. Recall that the same rate constant calculated *in situ* and *in vivo* may differ by more than an order of magnitude (Figures 2.5 and 2.6).

Rate constants also become viable candidates for contributing to the goal of mapping molecules produced in phenotypes back to their origins in the genome (Figure 7.27).

5.4.7 Provide Location Data

By treating enzyme densities as the unknowns in pairs of simultaneous equations, both equal and unequal distributions of enzyme densities can be detected *in situ* (Figure 2.2).

5.4.8 Supply Universal Data

A data type becomes universal when it can integrate data mathematically across the biology literature. Mathematical markers, connection ratios, ratio chains, biochemical densities, enzyme densities, membrane densities, and rate constants serve such a function by translating published data into complex data types.

5.4.9 Connect Data Across Multiple Levels of Complexity

When exploring morphological and biochemical data separately to look for reproducible patterns (level 1 complexity), large databases populated with data strings generated from thousands of papers are needed. When combined, however, morphological and biochemical data can provide ready access to biological rules at the level of individual papers (complexity levels 2 to 6).

5.4.9.1 Level 1 Complexity

The principle data types of level 1 complexity include mathematical markers and connection ratios arranged most often as triplets. Ratio chains (Equations 7.9 and 7.14), which connect biochemical or morphological data into linear strings, fall into this category.

5.4.9.2 Levels 2-6 Complexity

Units of enzyme activity and membrane surface areas (related to a gram of the parent structure) provide the raw data for calculating two complex data types: enzyme (U/S) and membrane (S/U) densities. Structure-function equations, which encapsulate biological rules, use these densities to predict unknown biochemical activities and membrane surface areas and to connect multiple levels of complexity.

5.4.10 Black Box Data

When first encountering the biology literature, the beginner soon becomes overwhelmed by the abundance of methods and data types. Under reductionism, we can put practically anything we wish into an experimental black box because its overriding goal is to generate changes that can be demonstrated statistically.

Under complexity theory, we replace the black box approach by including data types derived from biological rules:

(1) $x{:}\,y{:}\,z, \ldots, n\ \ (ratio)$,

(2) $Structure - Function\ Rule$

$f(x) = mx$

(3) $Structure - Function\ Change\ Rule$,

$f(x) = mx + b$, where

m equals an ED or MD, and

$ED = U/S\ \ (enzyme\ density)$

$MD = S/U\ (membrane\ density)$.

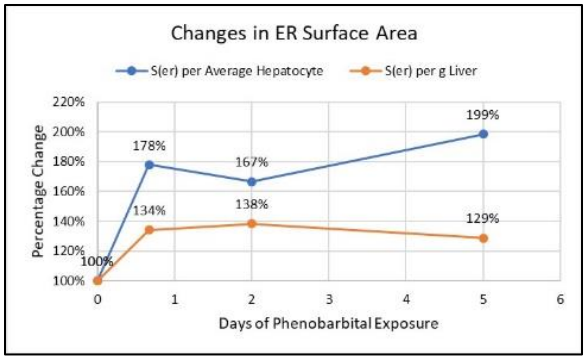
5.5 Biological Change – By Rule

If all biological data are related to something, then the question being asked of biology should be directed to that something capable of answering the question being posed. The point being that each of the many somethings supplies its own interpretation of the answer.

An organism, organ, cell compartment, average cell, one gram (or cm^3) of a tissue or organ, a mg of protein, et cetera all qualify as somethings. Although these data references are not equivalent, most are routinely used interchangeably to test biological data for significant differences with the belief that all such tests are detecting biological events. Since this is not the case, the biology literature is awash with contradictions. This is one of the important lessons that biology teaches with its complex data types.

5.5.1 Matching the Right Question to the Right Data

Experiment questions are highly dependent on data type. To get the right answer to our question, we must know what data type to ask. For example, let's ask a gram of liver and an average hepatocyte to tell us how much the ER membranes increases in response to the drug phenobarbital. Figure 5.4 illustrates the answers supplied by these two different data types.



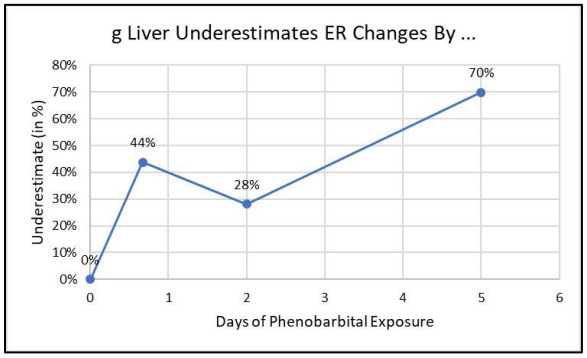


Figure 5.4 Top: The average hepatocyte displayed changes large enough to detect significant differences at all three time points with acceptable P values, whereas the gram of liver data didn't do as well. [In the original study a change was detected only at day 2 using a specific dimensions reference.] Bottom: On average, the gram of liver detected only 44% of the actual changes in the ER membranes. At day 5, however, it missed 70% of the change (Original data adapted from Stäubli et al., 1969).

The average hepatocyte delivered a substantial change (99%), whereas the gram of liver struggled to detect a change (29%). Figure 5.4 shows just how ineffective a gram of liver can be when used to detect a biological change. Missing 70% of the change at day 5 is difficult to write-off as an unavoidable methodological error.

Now consider Figure 4.26 in the previous chapter. What does it say about the membrane changes? May we conclude that the hepatocytes respond to phenobarbital by adding new ER membrane quickly and then enriching it progressively with new cytochrome P450 molecules? No (see Figure 5.4). We're asking a gram of liver to tell us something about the changes in ER membranes when we should be asking an average hepatocyte.

The point of Figure 4.26 was to ask the gram of liver to cancel out so we could replace it with a one square meter of membrane surface area – the reference for the enzyme densities (U/m^2). In turn, we could use the enzyme densities (expressed as a membrane density ($MD = ED^{-1}$)) with the total enzyme activity to calculate the relative average cell changes in ER surface areas, as shown in Figure 5.4.

5.5.2 Right Questions = Right Answers

Knowing how to ask biology questions becomes a key piece of our new and evolving skill set. We are learning what to ask and where and when to ask it.

Ask a gram (or cm^3) of an organ or tissue for:

1. Morphological values/g tissue
2. Biochemical values/g tissue
3. Ratios of parts (morphological or biochemical)
4. Enzyme densities
5. Membrane densities
6. Numbers of cells/g tissue

Ask an enzyme or membrane density to:

1. Access biological complexity
2. Link the events that define a change
3. Detect steady and transitional states
4. Provide rate constants
5. Look for reproducibility
6. Assemble mashups
7. Serve as a target for genotypic data

Ask an average cell for:

1. Ratios of parts
2. Changes in morphological components and biochemical constituents
3. Changes in cell volume

Ask an organ, gland, or tissue for the:

1. Ratios of parts
2. Total amounts of a cellular component (morphology)
3. Total amounts of a cellular constituent (biochemistry)
4. Changes in the number of cells
5. Relative changes in an average cell (when the total number of cells remains constant)

CHAPTER 6

DATABASES

Solving biology is largely an exercise in allowing two things to interact – research data and technology. Although solutions require unique data sets derived from databases, interpreting outcomes benefits importantly from hands on experience with laboratory methods, relational databases, programming, and standard software packages (e.g., PowerBuilder (Sybase), Excel and Access (Microsoft), Mathematica (Wolfram), and Photoshop (Adobe)).

6.1 Getting Started

6.1.1 Steep Learning Curve

The hardest and most challenging part of the discovery process consists of building an original database from the biology literature. The difficulty arises from the need to do several ill-defined things at the same time – without knowing at the outset what will eventually work. These include setting up a nomenclature (the names of parts preferred by authors), designing and testing database models (logical and physical), and programming user interfaces. The beginner and expert alike can expect a steep learning curve punctuated with more than a few false starts.

6.1.2 Access to Publications

The second most challenging task includes finding the best papers to put in the database and then securing copies. Experience suggests that at least three approaches work: request reprints from authors, download open access copies from the internet, and copy articles from journals. To end up with a production database containing data from 5 to 6,000 papers, expect to look at more than 10,000 articles.

6.1.3 Strategy

Databases become a primary problem-solving tool when dealing with biological complexity. Once stored in a database, published data can be translated into derived databases (parallel complexities) to solve specific problems. Solutions appear as patterns, graphs, and equations.

6.2 Stereology Literature Database

6.2.1 Database Models

Relational databases involve two models, logical and physical. One designs, the other executes.

6.2.1.1 Logical Model

Figure 6.1 displays the logical model used for the stereology literature database.

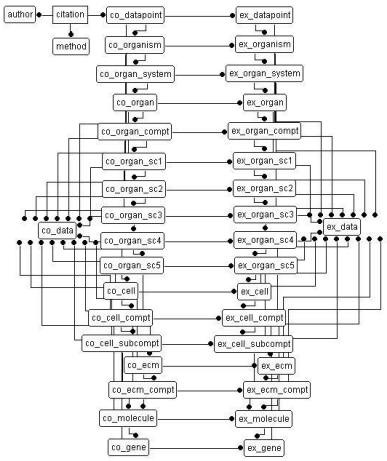
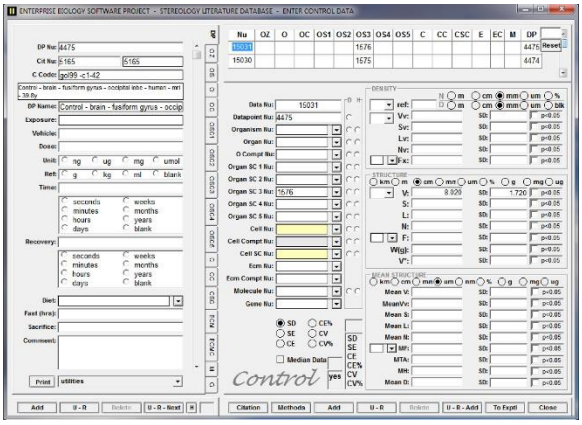


Figure 6.1 The logical database model for the biology literature includes a collection of entities (boxes) and relationships (lines), as defined by the rules of relational databases (From Bolender, 2001a).

The model includes the biological hierarchy (genes to organisms) in duplicate, with connections extending between control and experimental data. Although such a model increased the difficulty of data entry, it proved to be quite effective for finding clues and generating derivative databases.

6.2.1.2 Physical Model

The physical model includes the user interface where data are entered or retrieved (Figure 6.2).



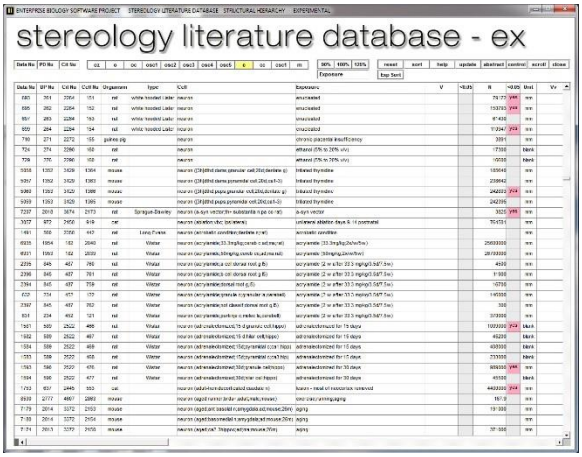


Figure 6.2 Top: The data entry process consists of assembling a hierarchy of parts (entities) by moving from one tab to the next (left) and then assigning numerical values to the parts (right) (From Bolender, 2001a). Bottom: The stereology literature can be reformatted into any number of different configurations. This table translates concentrations and absolute values into data pairs, designed specifically to hunt for evidence of reproducibility in the literature (From Bolender, 2001a).

6.2.2 Support Databases

Populating a literature database relies on other databases to help with data entry. The nomenclature and calculator databases serve as examples.

6.2.2.1 Nomenclature Database

The problem of assuring that the names in the database reflect those being used by most authors in the literature can be solved by compiling a database of lookup tables (Figure 6.3). By using these tables as templates, new entries adhere to the naming conventions preferred by most authors.

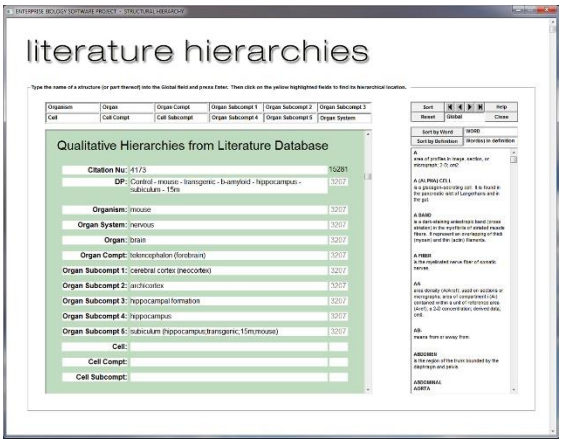


Figure 6.3 The task of standardizing data entry to a common set of terms and hierarchical locations requires a familiarity with the literature that comes only after entering data from thousands of papers. The result is a data entry format and nomenclature preferred by most authors. The green screen serves as the hierarchical template for data entry; terms and definitions appear at the right (From Bolender, 2001a).

6.2.2.2 Calculator Database

An unexpected finding was that a substantial proportion of published data appear exclusively in graphs. Consequently, a simple to use program was needed to speed the task of converting such graphs into numerical values. The data entry screen shown in Figure 6.4 offers an effective solution. A user supplies the information requested by the work screen and in turn the program calculates and displays the numerical value (highlighted in yellow).

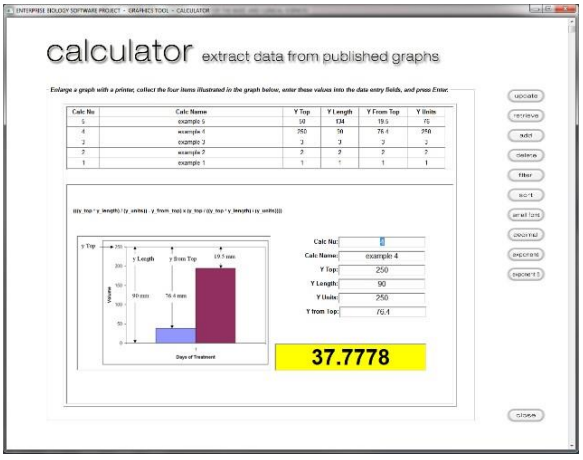


Figure 6.4 This work screen simplifies the task of turning graphs into numerical data (From Bolender, 2001a).

6.2.3 Derived Databases

Once designed and populated, the stereology database was used to generate derived databases. These databases, which usually consist of a single table, address specific issues. The most effective problem-solving format includes parallel complexities, which translate the original published data into patterns and data types that can address a wide range of questions.

6.2.3.1 Concentration Trap Database

The concentration trap database, which compares the same data expressed as concentrations and absolute values, identifies the risk involved when using concentrations to detect biological changes (Figure 6.5). With stereological data, for example, the database tells us that we can expect a correct result for concentration data only about 50% of the time.

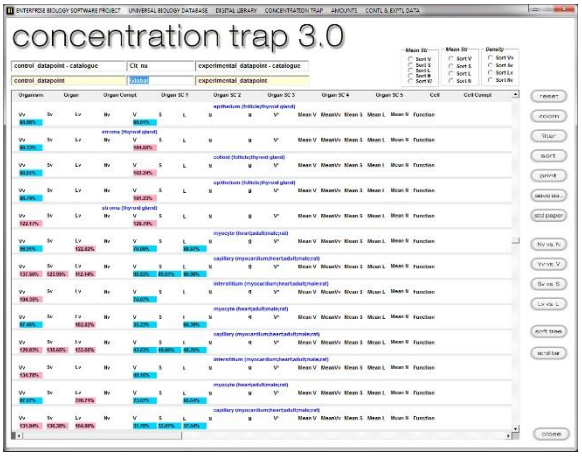


Figure 6.5 Concentrations and absolute values detect the same biological changes only about 50% of the time. Highlights identify the change: red = increase, blue = decrease, and green = no change. When – in the same row - the same highlight color appears on both the concentration and absolute values, they both detect a

change in the same direction – otherwise not (From Bolender, 2001a).

An important limitation of the concentration trap database is that neither of the two data types can be tested for accuracy. Since both concentration and absolute data carry methodological biases, the database offers keen insights into the problems created by operating within a theory structure based on reductionism. Suggested workarounds appear in the yearly reports and throughout the primer (Bolender, 2001a-2010).

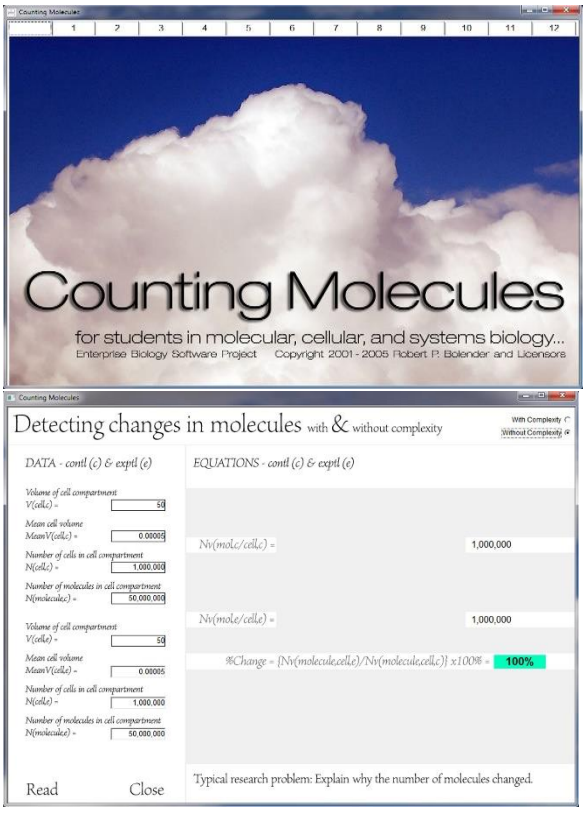
6.2.3.2 Counting Molecules Database

Solving biology mathematically benefits importantly from a basic understanding of the fine points surrounding biochemistry and intact tissue stereology. Since the results of each method involve many interacting variables, recognizing the contribution of each to the whole becomes a challenging task.

By taking biology apart, however, we introduce a new level of uncertainty. Now we must manage two entwined complexities simultaneously, one coming from biology and the other from our research methods (Bolender, 2016b). This requires a strategy based on minimizing methodological biases so that biological complexity can shine through.

Fortunately, simulators can provide a basic introduction to these biases by showing how key variables interact. The counting molecules program covers the essentials for both biochemistry and stereology. In Figure 6.6, panels 2 and 3 of the worked example show results typically published, whereas panel 4 explains what really happened. Comparing outcomes with and without complexity offers helpful insights into the problems created by trying to interpret biological data using the incomplete data sets available to most investigators.

After spending less than an hour exploring different scenarios with the simulators, the beginner begins to appreciate what can happen to experimental results when multiple variables are in play. The point to take from the exercise is that by keeping complexity out of our experimental approaches, we can't keep up with biology because it must operate as a complexity. Biology can't pick and choose the way we can because it must adhere to rules based on first principles.



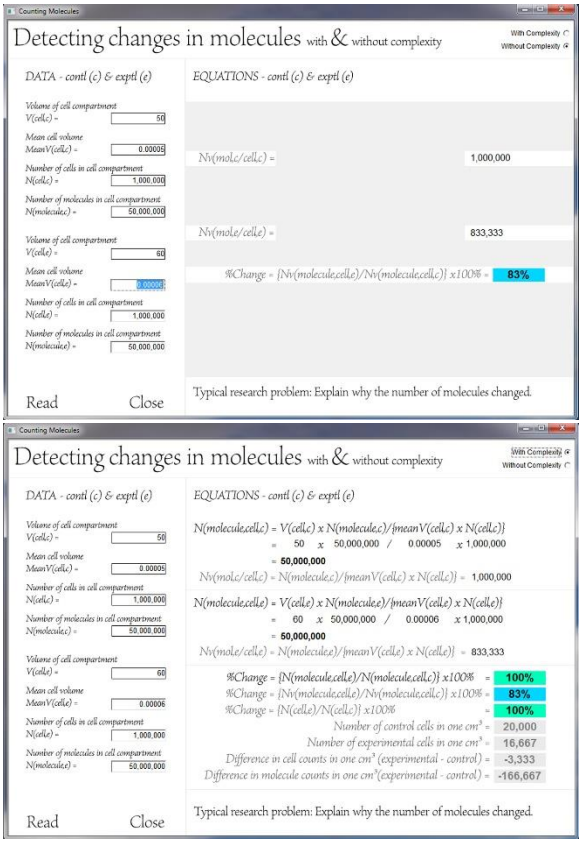


Figure 6.6 Detecting changes in molecules by comparing concentrations (i.e., optical densities) carries a risk. Panel 2 shows that the concentration of molecules – expressed as a numerical density (N/V) – is the same (100%) for both groups at the beginning of the experiment (1,000,000/cm³). After running the experiment, panel 3 shows a decrease in the number of molecules from the control value of 100% to 83%. By demonstrating a significant difference, most investigators would conclude that the experimental exposure effectively diminished this population of molecules. Many black box studies end here. Panel 4 tells a different story with the same data. It shows that the exposure caused the cells to swell slightly which meant that it took fewer of them to fill a cm³. Fewer cells meant fewer molecules per cm³. In fact, the absolute number of molecules remained unchanged. Hidden variables (panels 2 and 3) deprive the investigator of the critical information needed to interpret the results of this experiment correctly (From Bolender, 2005).

When detecting a biological change, the worked example shown in Figure 6.6 explains why estimates based on absolute data are preferable to those of concentration data. However, such a simulation doesn't account for the distortions in the reference volumes that can and do occur in experimental settings.

These disconnects between theory and practice occur whenever our experimental designs fail to account for the unique properties of biological systems and the biases introduced by our experimental procedures.

6.2.3.3 Design Code Equation Database

Design code equations enabled a fishing expedition, wherein stereological data were fitted to $R^2 \approx 1$ equations with the sole purpose of discovering if quantitative patterns (read biological rules) existed in published data. Since these equations found many corroborating patterns (e.g., Figures 6.7, 6.8), the next step consisted of reformatting the data into ratios to look for global patterns. To this end, the data populating the stereology literature database were converted into a database consisting of data pairs (axby), as described in Chapter 5.

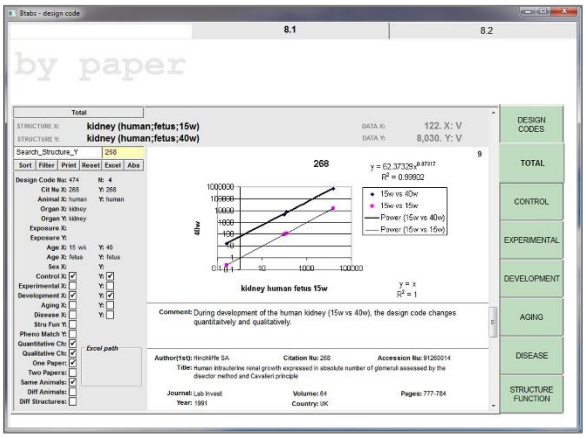


Figure 6.7 Local changes during development of the human kidney identify a distinct ordering of parts (Original data adapted from Hincliffe et al., 1991; From Bolender, 2003).

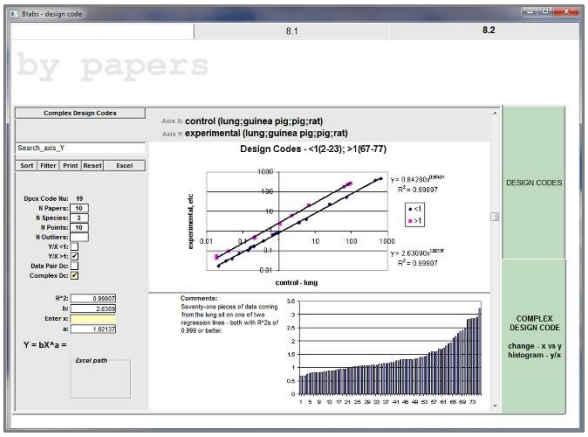


Figure 6.8 Global changes in the lungs of three animal species taken from 10 papers fit the same regression equations. (From Bolender, 2003). All the points included fall on the regression lines.

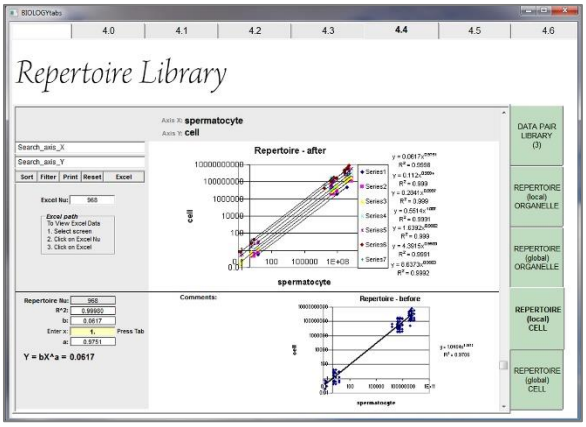


Figure 6.9 Regression analysis can turn clumps of data (before) into sets of parallel repertoire equations (after; From Bolender, 2004). In short, the different cell types in the testis show a remarkable degree of order.

6.2.3.4 Data Pair Ratio Databases

The $R^2 \approx 1$ (e.g., 0.999) equations can detect the relationship of one part to another as a ratio (x:y). A more efficient way of producing ratio-based equations (called repertoire equations) is to assemble a data set for each paper by generating all possible permutations - taking two parts at a time according to Equation 2.3. When applied to thousands of papers, this produces a universal biology database - a user-friendly version of the biology literature. By standardizing published data, an otherwise disconnected literature becomes one big experiment.

6.2.3.5 Universal Biology Databases

The data pair database yielded a wealth of new patterns, routinely finding order where previously none was known to exist. Figure 6.9 shows an example of repertoire equations turning clumps of data into an orderly set of equations wearing $R^2 = 0.999$.

By upgrading these equations to the decimal version (decimal repertoire equations), the sorting and filtering features of the new universal database simplified the task of finding global patterns and generalizations in the biology literature. The first major product of this decimal interpretation of the literature was the biology blueprint database.

6.2.3.6 Biology Blueprint Database

The purpose in assembling the blueprint database was to determine if similar patterns existed within and across species. It allows us to ask a fundamental question. To what extent do species known to share similar genes share similar phenotypes?

Assembling such a database consisted of first assigning data pairs – expressed as x:y ratios - to 81 bins (with intervals ranging from 0.001 to 100000) and then recording the frequency (n) of a given ratio in each bin (Figure 6.10).

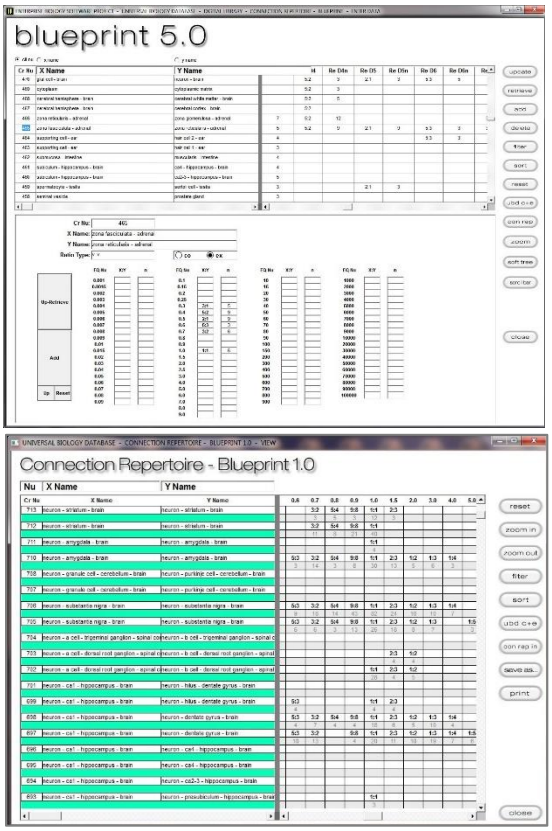


Figure 6.10 Top: Data entry consisted of linking all the connections (ratios) associated with a given pair of parts. Bottom: The biological blueprint documents the distribution of data pairs, ratios, valences, and frequencies (From Bolender, 2006).

The resulting blueprint shows how organisms define the relationship of one part to another with ratios and the extent to which a given phenotype displays a given relationship – within and across species.

In effect, the blueprint offers an empirical overview of the quantitative core of biology – as it exists in the postmortem data set of the stereology literature database. It shows that biological parts larger than molecules display a stoichiometry of whole numbers with many similarities existing within and across species. This provides preliminary evidence to suggest that organisms sharing the same parts with the same ratios are reading from the same genetic scripts. Moreover, the rules supervising these shared ratios appear to be highly conserved across biology. [Point: Although these patterns are reproducible, they are not necessarily accurate because stereological estimates typically carry methodological biases.]

The blueprint also serves as a convenient lookup table for finding specific phenotypes (Figure 6.11). For example, a given pair of parts (ax:by) often display several distinct valences, characterized as multiples of whole numbers (x:y). For example, the ratio of mitochondria to peroxisomes can be 10:1, 20:1, and 33:1 – depending on the cell, animal, and experimental setting. When writing simulations, reconstructing networks of equations, and trying to explain complex changes, such phenotypic information contributes importantly to the modelling process.

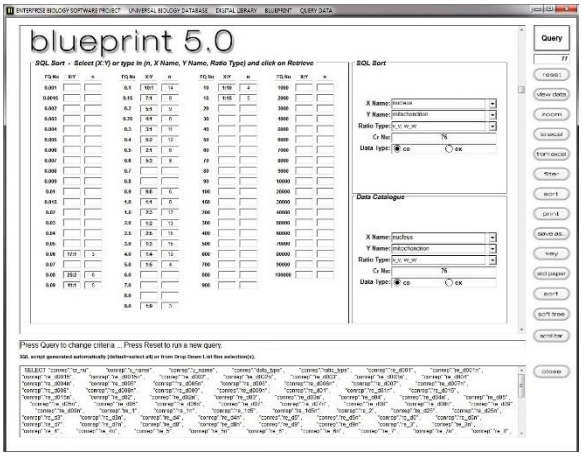


Figure 6.11 The SQL interface of the biology blueprint database shows the result of selecting an x:y ratio of 1:2. As items are selected from the query screen, the SQL script at the bottom of the screen updates accordingly. Clicking on the Query Button sends the request to the database, which promptly returns the information requested (From Bolender, 2006).

The blueprint database suggests that biology has evolved a universal parts inventory to draw from when assembling species, growing, making repairs, or adapting to disease. In effect, the widespread distribution of identical data pairs highlights this modular property of biology.

Figure 6.12 summarizes the blueprint modules. Roughly 40% of the total blueprint entries

(4,296) rely on just six decimal ratios (x:y): 50 to 1, 10 to 1, 3 to 1, 1 to 1, 2 to 3, and 1 to 10.

Decimal Repertoire Equation	Sum	%	Proportion (X:Y)
0.02	106	6.5	50 to 1
0.1	237	14.6	10 to 1
0.3	296	18.3	3 to 1
1.0	469	29	1 to 1
1.5	311	19	2 to 3
10	200	12	1 to 10

Figure 6.12 The figure lists the ratios in the biological blueprint occurring with the greatest frequency (From Bolender, 2006).

For neurons, however, the percentage of the most popular ratios goes up to about 70%. Neurons prefer four decimal ratios (x:y): 3 to 1, 2 to 1, 3 to 2, and 1 to 1.

The point to take from Figure 6.12 is that biology controls the relationship of one part to another deliberately. Since the same ratios can apply to parts ranging in size from small to large, it looks as if the entire biological hierarchy is subject to a common set of design rules. By simply forming ratios of small whole numbers, we can see the range of the rules biology uses to order its parts. Stoichiometry, a first principle well-known in chemistry, also seems to apply to biology. Either biology has come up with its own approach to forming ratios, which it may do genetically, or it may be piggybacking on the order already established by chemistry. Alternatively, it may be doing both.

6.2.3.7 Triplet Databases

Using ratio data offers several advantages not the least of which is the ability to increase our level of play. Starting with a relatively small number of published data points, we can end up with a considerably larger data set containing far more information.

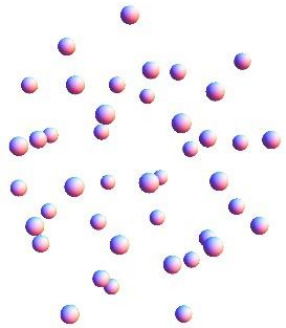
Consider data triplets – having three named parts a, b, and c with values x, y, and z. Three parts taken three at a time can be arranged six different ways: ax:by:cz, ax:cz:by, by:ax:cz, by:cz:ax, cz:ax:by, and cz:by:ax.

The reason for including six copies of the same information is that it optimizes outcomes. When looking for global patterns across many papers, we are least likely to miss a match when all possible permutations are included. Moreover, if a single triplet misses a pattern because of the decimal binning, we still have as many as five additional chances to catch it.

6.2.3.8 Mathematical Mapping Databases

The mathematical mapping databases were developed specifically to work with the graphing programs of Mathematica. The first such database, which included 2,000 triplet ratios, came from a single paper that used MRI to estimate the volumes of 42 parts in the brains of normal and schizophrenic patients (Goldstein, et al., 1999).

Figure 6.13 shows the original 42 brain parts first as a collection of isolated parts and then as a complexity displaying both parts and connections. Alternatively, we can display the connectivity map of the cerebral cortex in two dimensions (See Figure 3.10).



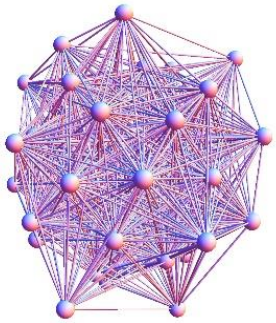


Figure 6.13 The original data set of the normal cerebral cortex included 42 isolated data points expressed as volumes. The 3D plot shows how 42 parts of the normal human cerebral cortex interconnect (Original data adapted from Goldstein, et al., 1999; From Bolender, 2011).

Let's look at the mapping procedure with an example. Consider two data pairs that relate one part to another and share a similar part. Parts a, b, and c with their ratios (1:2:3) generate, for example, the following data pairs:

$a1 : b2$
$a1 : c3$

This means that a1 maps to both b2 and to c2:

$a1 \rightarrow b2$
$a1 \rightarrow c3$, and
$b2 \rightarrow a1 \rightarrow c3$

When mapping parts of the brain with Mathematica, we would substitute the following notation:

"$frontal\ pole$" → "$cingulate\ gyrus$"
"$frontal\ pole$" → "$insula$" .

Such a mapping strategy provides a convenient way of visualizing large biological data sets before and after a change. For additional examples, see Bolender, 2011.

If we focus on the relationships of just one part – the angular gyrus – we can see that it connects to everything else quantitatively by sharing similar ratios (Figure 6.14). Notice in the figure that some parts (identified by blue dots) attract more connections than others. Recall that earlier we found a similar pattern for cell components (Figures 3.7 to 3.9).

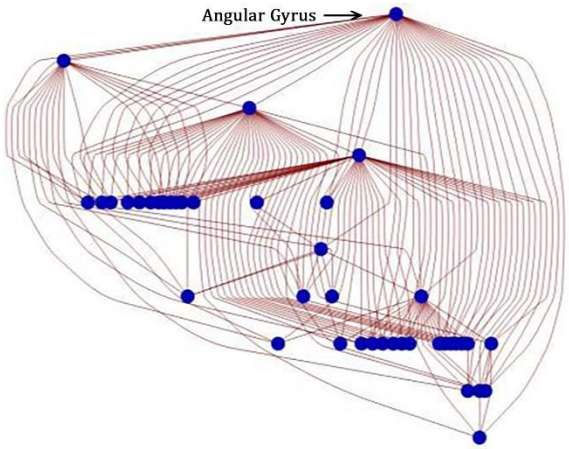


Figure 6.14 The mathematical map identifies the parts and connections of the angular gyrus in the normal human cerebral cortex. Notice that all 42 parts are connected (Original data adapted from Goldstein et al., 1999; From Bolender, 2011).

6.3 Internet Brain Volume Database (IBVD)

With the online publication of the Internet Brain Volume Database (IBVD; Kennedy, et al., 2012; Poline et al., 2012), we now have access to a data set well-suited to the task of looking for patterns and generalizations in health and disease. To this end, the published data were converted into mathematical markers and analyzed graphically with Mathematica.

This process includes several steps. After entering the names of the parts from a given paper as a string (a, b, c, … n) into Mathematica, the permutations program returns a list of all possible triplets - taking n parts three at a time (Equation 2.3). Next, the list of names is copied to an Excel spreadsheet where their numerical values are entered, used to calculate ratios, and assigned decimal repertoire values (Figure

7.13). The next step consists of copying all the data from individual papers to a single spreadsheet, saving it as a text file, and exporting the file into the table of a relational database (Access (Microsoft)). At this point, the database is sorted, filtered, and returned to Excel to be reformatted for analysis in Mathematica.

As a general observation, a spreadsheet works best for modifying data, whereas the database does extremely well at finding complex patterns especially in large data sets. See the yearly reports (Bolender, 2012 to 2015) for further details and worked examples.

6.3.1 Clinical Diagnosis

We have two general approaches to diagnosing disorders of the brain - subjective and objective. The subjective one – presents the greater challenge because extensive overlap exists between disorders and symptoms (Figure 6.15). Consequently, experts viewing the same patient do not always arrive at the same diagnosis. Such a reality confirms the wisdom of pursuing objective approaches to clinical diagnosis.

IMPAIRED IN DISORDER	ADHD	ALCOHOL	ALZHEIMER	ASPERGERS	AUTISM	BIPOLAR	BORDERLINE PD	MAJOR DD	OCD	PANIC DISORDER	PTSD	SCHIZOPHRENIA
PHYSICAL HEALTH		X						X	X	X	X	
SOCIAL FUNCTIONING	X	X	X	X	X	X	X	X				X
EMPLOYMENT-ECONOMIC	X	X	X	X	X	X	X	X	X	X	X	X
DELIQUENT BEHAVIOR		X				X	X					
SUBSTANCE ABUSE		X										
PHOBIA/PANIC/ OBSESSION									X	X	X	
NEGATIVE EMO-TION		X	X			X	X	X	X		X	
RISK OF SUICIDE		X					X	X		X		X
EXCITEMENT/ ELATION						X						
HYPERACTIVITY	X					X						
REALITY		X				X		X				X
TRUST												X
FRIENDLINESS				X	X							X
JUSTICE		X	X									
WISDOM	X	X	X	X	X	X	X	X				X
CAUTION	X					X	X					
STABILITY			X			X	X					
CONSCIENTIOUS-NESS							X		X			X
CONFIDENCE						X	X	X				
INDEPENDENCE			X			X	X			X		
OVERALL FUNCTIONING	X			X	X	X	X	X	X	X	X	X

Figure 6.15 The table identifies symptoms for various disorders as impairments. Given their subjective nature and the fact that the same impairment often applies to many disorders, making a differential diagnosis becomes problematic (Adapted from Internet Mental Health © 1995-2015 Phillip W. Long, M.D.; From Bolender, 2015).

Since the IBVD includes clinical data from normal individuals and those with diagnosed disorders, an objective approach now becomes a viable option. By translating the MRI data of the IBVD into mathematical markers, we can generate databases containing a large enough number of patterns to work out an objective approach to clinical diagnosis. Such diagnostic databases come with the added advantage of showing us just how tightly biology connects and controls everything by rule.

6.3.2 Diagnosis Database (Unique Triplets)

The first diagnostic database includes triplet mathematical markers wherein each marker is unique to a given disorder of the brain, of which

there are 26. Figure 6.16 illustrates 9 of the 26 disorders. The central point of each circular radiation identifies a disorder, which is surrounded by a set of unique markers. The number of markers in a set ranges from three to tens of thousands.

Within this closed environment (called a data cage), we can be confident that any mathematical marker taken from the diagnosis database, relabeled as <unknown>, and run against the database will always give the correct diagnosis (recall Figure 3.23). Why? Because in a "caged" database populated with mathematical markers unique to each disorder, false negatives and false positives do not exist.

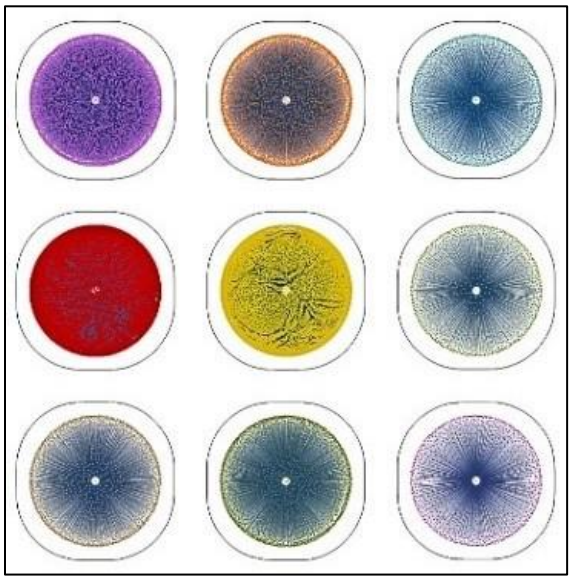


Figure 6.16 A data cage designed for the human brain includes a collection of twenty-six disorders (only 9 shown here) – each of which exists in the figure as a central point surrounded by a radial set of unique mathematical markers (Original data adapted from Kennedy et al., 2012; Adapted from Bolender, 2015).

6.3.3 Diagnosis Database (Unique Quadruplets)

The second diagnostic database replaced the triplet markers (ax:by:cz) with quadruplets (ax:by:cz:dq). It produced the same result (see Figure 3.21) but used a larger number of markers (3,651,770).

What do these diagnostic databases tell us? We can identify disorders of the brain – quickly and easily - by translating clinical data into mathematical markers and then use the resulting databases to look for objective patterns (Figures 3.22 and 3.23; Bolender, 2008).

The framework of the relational database does all the work for us. It vets the data for uniqueness and can run diagnostic tests by challenging the database with unknown markers. Both steps involve running the "find duplicates" routine (a SQL script) built into Microsoft's Access Database. Details and worked examples appear elsewhere (Bolender, 2015; Appendix II)

What are we learning? It appears that one of the secrets to working in a complex data environment, such as those of biology, is to design outcomes as equations with $R^2 \approx 1$ or to set the confidence level of predictions (e.g., diagnosis) at 100%. In short, we are learning to use biology's rules to solve our problems.

6.3.4 Diagnosis Database (Shared Triplets)

Recall that addressing a complex problem typically requires reformatting a preexisting database as a parallel complexity designed specifically to answer the question(s) being posed.

Let's say we want to study several disorders of the brain as a group, instead of just individual disorders. If we go back to the original diagnosis database, we can filter its contents such that we keep only those markers that have duplicates in at least three publications.

Shared markers can tell us two things. First, the patterns created by the mathematical markers tell us something about how biology produces disorders (by identifying duplicated modules) and second, how closely one disorder resembles another (by applying cluster analysis). In short, duplicate markers allow us to probe the phenotype for the causative factors underlying the disease process.

To get a global view of the disease process, we can translate the MRI data of the IBVD into a composite brain carrying duplicate markers for 21 different disorders (see Figure 3.24). Next, by parsing this brain stepwise with the CommunityGraphPlot from Mathematica, we can assemble a relationship tree that identifies families of closely related disorders (Figure 6.17).

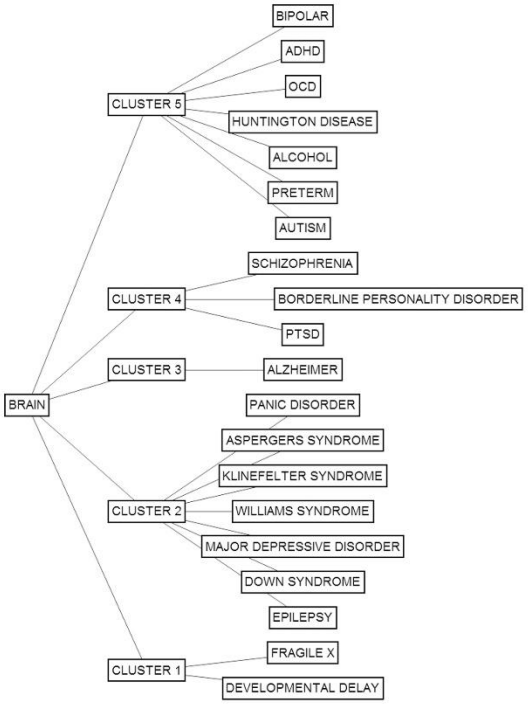


Figure 6.17 The graph shows the relationship of clusters to disorders in the composite human brain, unfolded as a function of shared mathematical markers (From Bolender, 2015).

Figure 6.18 compares two approaches to diagnosing disorders of the brain. The top panel illustrates the subjective approach, whereas the bottom one does it objectively with a database containing mathematical markers. Notice that the borderline personality disorder shares a sizable portion of its markers with schizophrenia – the pattern displayed by both plots. The clue that triggered this quantitative solution to the diagnosis problem came from Figures 3.14 and 3.15.

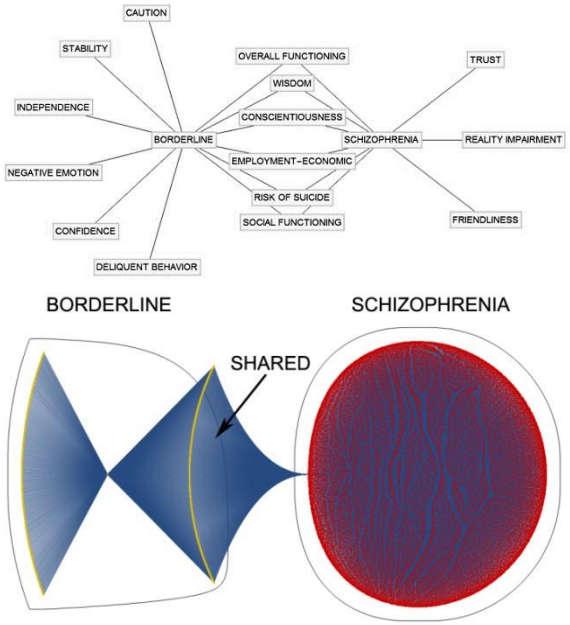


Figure 6.18 Graph and community plots illustrate the extent to which different disorders share similar symptoms (Top: subjective) and mathematical markers (Bottom: objective); (Original data adapted from Kennedy et al., 2012; From Bolender, 2015).

Results collected with the shared triplets database point to a modular design as the common thread running through all the disorders. Operationally, biology appears to have access to a vast collection of building blocks that it uses to assemble normal and abnormal brains.

How is this helpful? If biology is changing patterns from normal to abnormal and we can detect such patterns with mathematical markers, then we know where to begin looking for the underlying causes. At some point, it might become possible to map shared

symptoms to shared mathematical markers (e.g., Figure 6.18).

6.3.5 Diagnosis Database (Shared Quadruplets)

Figure 6.19 summarizes the frequency distributions of duplicates for quadruplet markers. Even when the alphanumeric string of the quadruplet marker contains 8 variables (ax:by:cz:dq), 21% of the markers in the quadruplets database formed duplicates globally. This percentage amounts to more than two million diagnostic markers generated from the IBVD. Does this mean that 21% of the published data are reproducible?

Duplicates	Normal		Disease		Normal	Disease
	Total	Groups	Total	Groups		
2	832246	416123	1447108	723554	91.47%	76.45%
3	56139	18713	319635	106545	6.17%	16.89%
4	14616	3654	114984	28746	1.61%	6.07%
5	3660	732	8135	1627	0.40%	0.43%
6	1368	228	1722	287	0.15%	0.09%
7	462	66	840	120	0.05%	0.04%
8	768	96	288	36	0.08%	0.02%
9	432	48	270	30	0.05%	0.01%
10	60	6	0	0	0.01%	0.00%
11	66	6	0	0	0.01%	0.00%

Figure 6.19 The distributions of quadruplet markers suggest that the brain responds to the disease process by changing its connectivity. In disease, the percentage of markers tends to shift from 2 copies per group to 3 and 4. Of the 13,360,056 quadruplet markers, 2,802,799 (21%) were duplicates (Original data adapted from Kennedy et al., 2012; From Bolender, 2014).

Notice in Figure 6.19 that the percentage columns at the right identify a prominent shift in the frequency distribution of the duplicate groups from 2 copies to 3 and 4 copies. This suggests that the disease process brings about an **increase** in connectivity. [Recall the local finding of Figure 3.8.]

By comparison, the triplet database included 381,476 duplicate markers (Bolender, 2014) - , representing 47.2% of the total population. A database derived from the biology literature that contains as many as 64 copies of the same mathematical marker offers a compelling argument for the presence of reproducibility in biology and in published research data. At least the possibility now exists that we may be better at our job than some might lead us to believe (Baker, 2016; Pellizzari et al., 2017).

CHAPTER 7

CALCULATIONS

Solving biology includes a team effort involving data, rules, mathematics, and technology. Typically, calculations span several steps and technologies, which include databases, spreadsheets, equations, programming, and software packages. The goal of this chapter is to illustrate the range of the calculations used to extract and reinterpret published data. Worked examples - explained in greater detail - accompany the yearly reports and can be found online.

7.1 Collecting Data

7.1.1 Calculator Program

Solving biology begins with the biology literature. After identifying a paper of interest, the first step consists of organizing the data into a table that corresponds to the work flow of the data entry screens in the database. This speeds data entry and minimizes errors. Unfortunately, many authors prefer graphs over data tables when reporting their results.

Herein lies a problem. Graphs come in many different forms and figuring out how to extract numbers from points and lines quickly becomes a frustrating task. A workable solution to this problem appears in Figure 7.1, wherein a program offers a general solution to the data conversion problem. By following the directions given on the calculator screen, the data conversion process becomes a routine operation. [A helpful hint: An enlarged photocopy of a published graph makes it easier to measure distances and to keep track of the numerical values.]

The equation for making the conversion is as follows:

$$y\ value = (((y\_top \times y\_length)/ (y\_units)) - y\_from\_top) \times (y\_top/ ((y\_top \times y\_length)/(y\_units\ )))) \quad (7.1)$$

For example, when:

$y_{top} = 250\ mm$
$y_{length} = 90\ mm$
$y_{\_units} = 250\ mm$
$y_{\_from\_top} = 76.4\ mm$

$$\boldsymbol{y\ value = 37.7778}$$

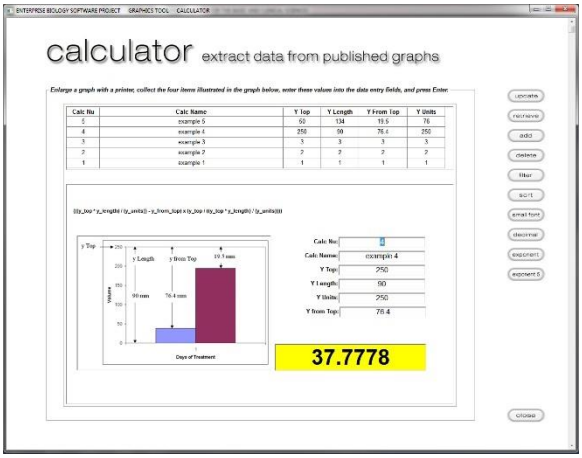


Figure 7.1 Since many publications report data exclusively as graphics, a work screen simplifies the task of translating graphical data back into numerical values (From Bolender, 2001a).

7.2 LEVEL 1 – PATTERNS

7.2.1 CONCENTRATION TRAP

It's easy to imagine that once we acknowledge something to be true in one setting, that it must also be true in other settings. Sometimes this works, other times not. Consider, for example, chemistry and biochemistry. Although it's perfectly acceptable in chemistry to detect changes in molecules by comparing concentrations, the same permission does not extend to biochemistry when our goal is to interpret molecular changes as biological changes.

Why? A chemical change uses a constant data reference (e.g., ***1 ml of volume***) to detect a change, whereas a biochemical change related to cells uses a variable data reference (e.g. ***the contents of 1 ml of volume***) to detect a change. This means that each type of chemistry comes with its own set of rules. Ignoring such rules, which happens routinely, introduces chaos into the literature and leads to unfortunate consequences such as the ongoing reproducibility crisis.

Consider the challenge we face. In designing an experiment in biology, we adhere to a simplified methodology (reductionism), but interpreting results requires that we evaluate our findings within the complex reality of biology. In effect, we find ourselves operating within a research model filled with inconsistencies. The unfortunate consequence of this mismatch between theory and practice is that we unwittingly fall into traps of interpretation.

All too often, a change in the same part (A) gives one result as a concentration (A/reference volume), but quite a different one when expressed as an absolute value (A). Let's see how and why this happens.

Concentration Data

$$control(co) = A_{co}/V_{co\_ref} \quad (7.2)$$

$$exptl(ex) = A_{ex}/V_{ex\_ref} \quad (7.3)$$

Change in Concentration

$$\frac{exptl(ex)}{control(co)} = (A_{ex}/V_{ex_{ref}})/(A_{co}/V_{co_{ref}}) \quad (7.4)$$

Outcome 1: If $V_{ex\_ref} = V_{co\_ref}$, the reference volumes cancel out and the result is interpretable:

$$\frac{exptl(ex)}{control(co)} = (A_{ex})/(A_{co}) \quad (7.5)$$

Outcome 2: If $V_{ex\_ref} \neq V_{co\_ref}$ the reference volumes do not cancel out and the result is uninterpretable.

$$\frac{exptl(ex)}{control(co)} = (A_{ex}/V_{ex_{ref}})/(A_{co}/V_{co_{ref}}) \quad (7.6)$$

How often does ***Outcome 1*** occur? According to the concentration trap database (Figure 6.5), concentration data (***Outcome 1***) agree with their absolute data counterparts only about 50% of the time. This tells us that that we can expect morphological and biochemical data to fall routinely into concentration traps. Indeed, they do (Figures 6.5 and 7.30)

The key to getting our arms around a biological complexity such as a change begins with the understanding that biological variables (read parts and connections) are interconnected and interdependent. In biology, nothing acts alone. Simulators can help to explain why this is the case.

7.2.2 BLACK BOX SIMULATOR

By taking biology apart to study it, we not only dumb-down biology by destroying its complexity, but we're also cutting off our access to biology's rules. This makes it difficult to play biology's game and even harder to win.

Let's begin by considering the sources of the contradictory results detected by the concentration trap by identifying some of the rules in play. To this end, we can return to the calculation work screens of the *Counting Molecules* simulator (Figure 7.2) introduced earlier (Figure 6.6).

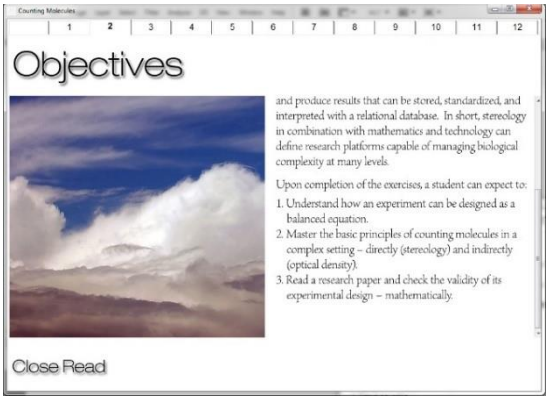


Figure 7.2 The Counting Molecules simulator introduces the beginner to the risks of interpreting biological data taken out of context (From Bolender, 2005).

Next, let's do an experiment using the standard black box approach wherein we compare the concentrations of a molecule estimated in an organ before (contl) and after (exptl) exposure to a hypothetical toxin. We can see in Figure 7.3 that the concentration of the molecules increased by 39%, a change large enough to detect as a significant difference. Most readers seeing such a result would support the author's conclusion that the number of molecules increased.

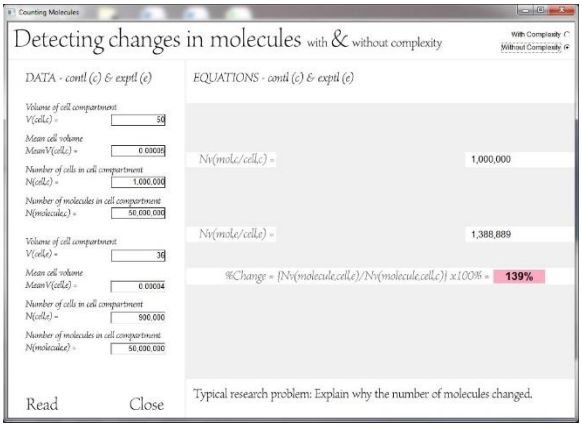


Figure 7.3 A biological change is the product of many interacting variables, not just one. By comparing just concentrations, detecting a biological change has roughly a 50-50 chance of being correct (From Bolender, 2005).

Given the black box design of the experiment in Figure 7.3, can we safely conclude that the toxin increased the number of molecules? No. Did the study detect an increase in the concentration of the molecules? Yes. Did the number of molecules increase? No.

What happened? When we open the black box (shown in Figure 7.3 as a gray overlay) and look inside, we can see what our imaginary toxin did to the cells (Figure 7.4).

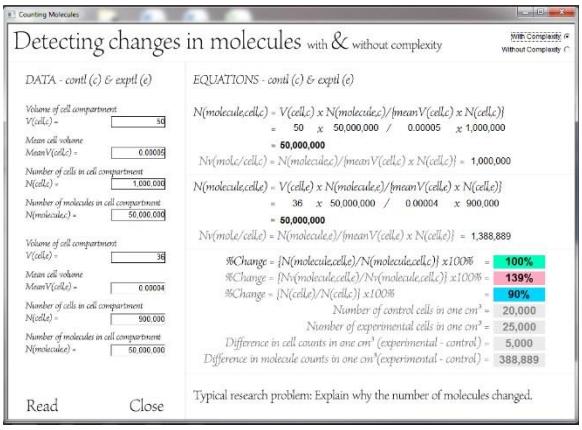


Figure 7.4 Molecular changes embedded in biological complexity. Highlights identify the change: red = increase, blue = decrease, and green = no change. From Bolender, 2005).

The toxin:

- Had no effect on the number of molecules in an average cell.
- Killed 100,000 cells.
- Decreased the average cell volume by 20%.
- Moved 5,000 extra cells into a cm^3 of cells.
- Moved 38,889 extra molecules to a cm^3 of cells.
- Increased the concentration of the molecules in a cm^3 of cells by 39%.

What do the results of the simulation in Figures 7.3 and 7.4 tell us? It's often impossible to interpret biochemical results correctly without including complexity. The concentration data (without complexity) detected a 39% increase in the number of molecules that didn't happen and missed the fact that the toxin killed 100,000 cells per cm^3 of tissue. In short, an over-simplified approach to biology can be hazardous to both cells and researchers.

Unfortunately, the results shown in Figures 7.3 and 7.4 only begin to hint at the calamities put into play by ignoring the complexity of a biological change. A simulator helps us to understand the nature of complexity by keeping track of multiple variables and showing us what can happen as we vary one or more of the variables at a time.

Avoiding complexity puts the very science of biology at risk. Detecting a biological change correctly is so tricky that the results of only 20-30% of refereed publications in the biology literature are now believed to be correct (e.g., Ioannidis, 2005).

7.2.3 Repertoire Equations

A repertoire equation defines the relationship of two sets of parts (x, y) such that the slope (b) and the R^2 of a power curve $(y = ax^b)$ both approach 1.0. Recall that a power curve with a slope of 1 becomes a linear curve.

Figure 7.5 plots organelles against the endoplasmic reticulum (er) with data coming from 16 papers. The resulting repertoire equation is given as:

$Y = 0.4549X^{1.00156}$, with an $R^2 = 0.999$. (7.7)

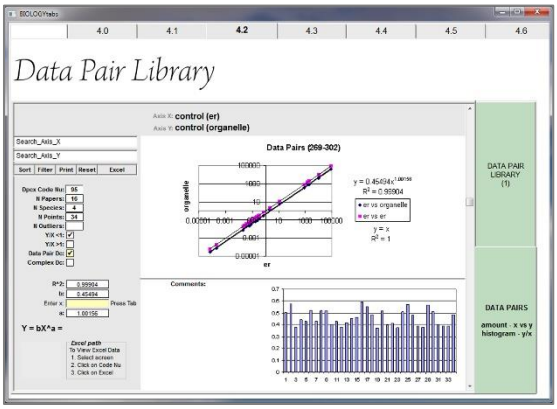


Figure 7.5 A high degree of order can exist in the relationship of the endoplasmic reticulum to other cell organelles. The figure shows that diverse organelles coming from different publications can occur in the same proportion with the endoplasmic reticulum (From Bolender, 2004).

Recall that repertoire equations provided an important early clue by showing that global order could be found in the biology literature. They did this largely by detecting an underlying order defined by ratios.

7.2.4 Decimal Repertoire Equations

Although repertoire equations (derived from data pair ratios) uncovered previously undetectable patterns, they became increasing unwieldy because they lacked distinct boundaries between adjacent patterns. This limitation created challenges of interpretation when filtering and sorting large database tables.

The problem was resolved by assigning the data pair ratios to distinct decimal bins, which were determined by calculating regression equations

for the data contained within each bin. The decimal steps were chosen such that the regressions predicted the published values with a maximum error of no more than ±15% - the typical error associated with stereological estimates. The decimal steps ranged from 0.0001 to 100,000. By switching to a decimal format, the data pairs (and subsequently triplets and quadruplets) displayed distinct boundaries, which facilitated the filtering, sorting, and pattern recognition routines.

The universal biology database table in Figure 7.6 lists a collection of data pairs along with a measure of the effectiveness of the repertoire equations in predicting the original values expressed as a percentage. When we compare the observed values for y to the expected values, most closely agree with only about 50 (0.1%) of the 50,000 predictions falling outside the ±15% interval.

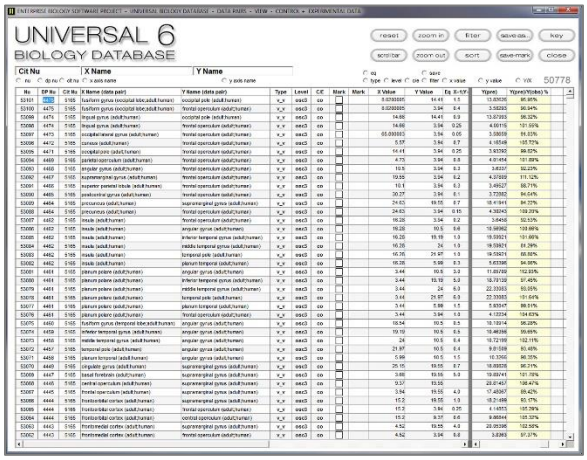


Figure 7.6 The data pair table includes ratios (y/x), ratios, repertoire equations, and an assessment of how close the equations predict the original ratios (see the yellow column); From Bolender, 2005.

By attaching each data pair ratio (y/x) to a decimal repertoire equation, for example, patterns previously hidden in amorphous clouds of data become detectable. Figure 7.7, for example, shows how a cloud of points (Before) quickly resolves into 14 decimal repertoire equations with R^2s > 0.9 (After).

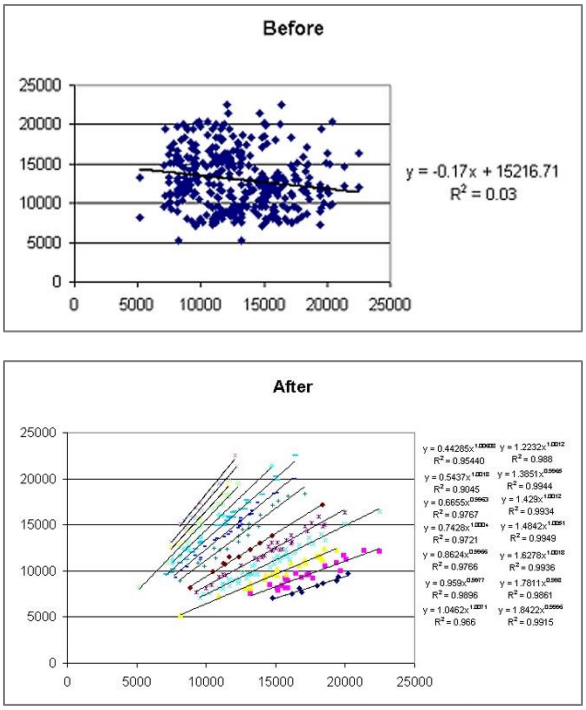


Figure 7.7 Before: Cell counts plotted from the lateral geniculate nucleus of the brain showed a single data cloud with little indication of order. After: The data cloud unfolded into 14 decimal repertoire equations with R^2s > 0.9. Decimal repertoire equations proved to be effective in distinguishing between closely related phenotypes quantitatively. Notice how the curves all point toward the origin, which tells us that biology likes to use the same ratios repeatedly (Original data adapted from Seecharan et al. 2003; From Bolender, 2005).

Figure 7.7 offers a key insight. In 58 isogenic strains of mice, the equations show that there are at least 14 ways to design a lateral geniculate nucleus – given the frequencies of three cell types (neurons, glial cells, and endothelial cells).

What does this tell us? Since most of the genetic modifications in the Seecharan study did not target genes related to the nervous system, these results suggest that making changes to the DNA - anywhere in the genome - can have unintended consequences at the level of the organism.

With genomes routinely being modified, we can expect to see such patterns of unintended consequences occurring elsewhere. Changing

battle-tested sequences in the genome – in the absence of broadly-based feedback from the phenotype – may end up producing more problems than it solves.

7.2.5 Growth Patterns

Patterns often undetected with standard methods of analysis become visible when expressed as decimal repertoire values (DRV). Figure 7.8, for example, uses such values to chart the development of the adrenal gland in the rat. Notice that the same patterns in the ratio of parts (yellow=yellow, green=green) appear, disappear, and reappear as the adrenal develops.

	28 Days		35 Days		42 Days		49 Days		56 Days		63 Days		70 Days		77 Days	
DRV																
0.2			1	16.67%			1	16.67%	1	16.67%	1	16.67%			1	16.67%
0.25	1	16.67%		0.00%	1	16.67%		0.00%		0.00%		0.00%	1	16.67%		0.00%
0.3		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
0.4		0.00%	2	33.33%	1	16.67%	1	16.67%	1	16.67%	2	33.33%	1	16.67%	2	33.33%
0.5	2	33.33%		0.00%		0.00%	1	16.67%	1	16.67%		0.00%		0.00%		0.00%
0.6		0.00%		0.00%	1	16.67%		0.00%		0.00%		0.00%		0.00%		0.00%
0.7		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%	1	16.67%		0.00%
0.8		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
0.9		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
1		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%	1	16.67%		0.00%
1.5	2	33.33%		0.00%	1	16.67%	1	16.67%	1	16.67%		0.00%		0.00%		0.00%
2		0.00%	2	33.33%	1	16.67%	1	16.67%	1	16.67%	2	33.33%	1	16.67%	2	33.33%
2.5		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
3	1	16.67%		0.00%	1	16.67%		0.00%		0.00%		0.00%	1	16.67%		0.00%
4		0.00%	1	16.67%		0.00%	1	16.67%	1	16.67%	1	16.67%		0.00%	1	16.67%
5		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
6		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%
7		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%		0.00%

Figure 7.8 Development of the adrenal gland displays repeating patterns (yellow, green) over time (left to right) (Original data adapted from Nikicicz et al., 1984; From Bolender, 2008).

Figure 7.9, which shows the effect of ACTH on the development of the hamster adrenal, also detects repeating patterns (yellow = yellow, green = green); DRV = decimal repertoire value.

Days	0		7		14		21		28		35	
DRV												
0.2	1	17%										
0.25		0%	1	17%			1	17%	1	17%	2	33%
0.3		0%	1	17%	2	33%	1	17%	1	17%		0%
0.4	1	17%		0%		0%		0%		0%		0%
0.5	1	17%		0%		0%		0%		0%		0%
0.6		0%		0%		0%	1	17%		0%		0%
0.7		0%	1	17%		0%		0%	1	17%		0%
0.8		0%		0%		0%		0%		0%		0%
0.9		0%		0%		0%		0%		0%	1	17%
1		0%	1	17%	2	33%	1	17%	1	17%	1	17%
1.5	1	17%		0%		0%		0%		0%		0%
2	1	17%		0%		0%		0%		0%		0%
2.5		0%	1	17%		0%	1	17%	2	33%		0%
3		0%	1	17%	2	33%	1	17%			2	33%
4	1	17%										

*Figure 7.9 Notice that the response of the adrenal to ACTH during development once again shows repeating patterns (yellow, green) over time (*Original data a*dapted from Malendowicz, 1986; From Bolender, 2008).*

In both cases, ratios of parts, but not the individual parts, detected the repeating growth patterns.

7.2.6 Triplet Markers (Worked Example)

Recall that mathematical markers represent an alpha-numeric string (ax:by:cz), which, for example, can consist of three named parts (e.g., a, b, c) with their corresponding numerical values given as ratios (x:y:z). The names may refer to any biological part and the numerical values typically characterize a volume, surface, length, or number. For the following worked example, the data set will consist of four named parts (a to d) and four volumes (5 to 20) – as shown in Figure 7.10.

Part	Volume
a	5
b	9
c	12
d	20

Figure 7.10 The data set associates 4 parts with 4 volumes.

The permutations program of Mathematica takes four parts (a, b, c, d) three at a time {3} to generate twenty-four triplets (Figure 7.11).

```
In[5]:= Permutations[{a, b, c, d}, {3}]
```

Out[5]=

a	b	c
a	b	d
a	c	b
a	c	d
a	d	b
a	d	c
b	a	c
b	a	d
b	c	a
b	c	d
b	d	a
b	d	c
c	a	b
c	a	d
c	b	a
c	b	d
c	d	a
c	d	b
d	a	b
d	a	c
d	b	a
d	b	c
d	c	a
d	c	b

Figure 7.11 Four parts taken 3 at a time yield 24 triplets ($4 \times 3 \times 2 = 24$).

In the next step (Figure 7.12), we copy these letter triplets and paste them into an Excel calculation template (columns D, E, F). In turn, we copy their numerical values from the Excel worksheet shown at the top to the one below: column B (**top)** to column H (**bottom)** as shown. Since column D contains the same names as columns E and F, the numerical values are also the same, but in a different order. By copying column x and storing it on an adjacent worksheet, we can fill columns I and J by copying and pasting all the stored values for a given column together. To do this, highlight the entire bottom screen and select custom sort. Next, sort (ascending) on column E, copy the stored values, and paste them in column I. The procedure is repeated for columns F and J. Notice that the repertoire ratios (columns K, L, M) appear automatically in the Excel template after entering the volumes. To view the calculation in the heading of the worksheet, highlight a value in the table.

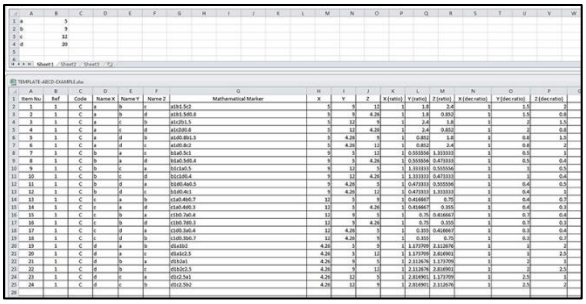


Figure 7.12 The two worksheets summarize the process of translating parts and ratios into triplet mathematical markers. Enlarge as needed. Such template worksheets are available online.

The last step consists of translating the repertoire values (raw data) into decimal repertoire values using a lookup table (Figure 7.13) and the raw ratios listed in columns x, y, and z (Figure 7.12).

DECIMAL REPERTOIRE VALUES

0.00001	0.04	60
0.000015	0.05	70
0.00002	0.06	80
0.00003	0.07	90
0.00004	0.08	100
0.00005	0.09	150
0.00006	0.1	200
0.00007	0.15	300
0.00008	0.2	400
0.00009	0.25	500
0.0001	0.3	600
0.00015	0.4	700
0.0002	0.5	800
0.0003	0.6	900
0.0004	0.7	1000
0.0005	0.8	2000
0.0006	0.9	3000
0.0007	1	4000
0.0008	1.5	5000
0.0009	2	6000
0.001	2.5	7000
0.0015	3	8000
0.002	4	9000
0.003	5	10000
0.004	6	20000
0.005	7	30000
0.006	8	40000
0.007	9	50000
0.008	10	60000
0.009	15	70000
0.01	20	80000
0.015	30	90000
0.02	40	100000
0.03	50	

Figure 7.13 Decimal repertoire values use 101 bins to store the published data (From Bolender, 2017). A value (x), for example, would be assigned to the 5 bin when x ≥ 5 and < 6.

This concluding step of data entry begins by sorting (the entire table) on the L column and entering the decimal values into column O (see Figure 7.12). When finished, column O is copied and saved on an adjacent worksheet. Finally,

sort the entire table on column M, copy the saved data of column O, and paste it into column P.

Why use such a complicated procedure? Since a paper with 15 parts generates 40,950 markers for a given time point, using the copying and pasting routines speeds the process and minimizes errors.

A final note. When dealing with large data sets in Excel (>1.2x10^6 rows), generating the triplets or quadruplets requires shuttling data between various programs and figuring out how to get the programs to do what needs to be done. The appendix attached to the 2016 report (Bolender, 2016a) includes a worked example with additional suggestions that may be helpful.

Figure 7.14, for example, shows a calculation template designed to generate triplet markers. To make it legible, the template is displayed in two parts – left (top) and right (bottom).

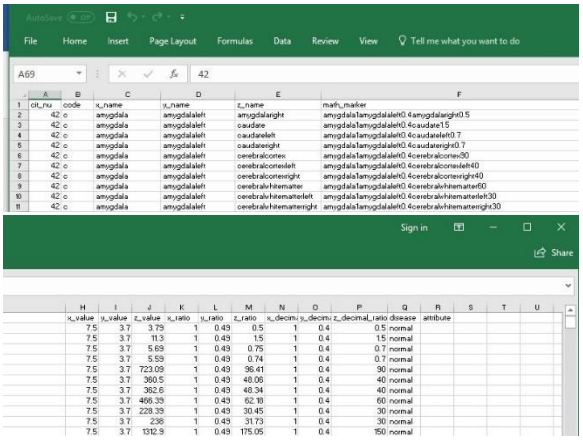
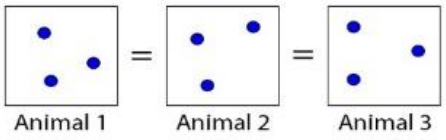


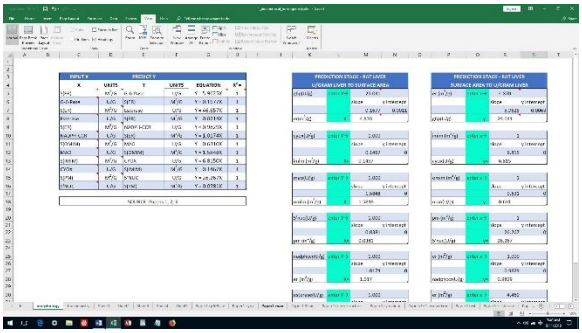
Figure 7.14 Triplets were produced using a worksheet template. The concatenate function - CONCATENATE (C2,N2,D2,O2,E2,P2) generates the mathematical markers automatically. The lookup sheet (Figure 7.13) was used to translate the repertoire ratios (K,L,M) into decimal ratios (N,O,P); (From Bolender, 2017).

In effect, this template-based procedure speeds the task of translating the data of standalone research papers into a single, universal biology database populated with highly interactive mathematical markers.

7.3 Level 2 – Biochemical Homogeneity



Biochemical homogeneity, which defines a relationship of structure to function mathematically, allows us to predict morphology from biochemistry and vice versa. Calculation modules, such as those shown in Figure 7.15, predict both morphology and biochemistry. The file illustrated in the figure (1_biochem-ical_homogeneity.xlsx) is available online as an attachment to the 2017 progress report.



INPUT X		PREDICT Y			
X	UNITS	Y	UNITS	EQUATION	R^2 =
S(ER)	M^2/G	G-6-Pase	U/G	Y = 5.9625X	1
G-6-Pase	U/G	S(ER)	M^2/G	Y = 0.1677X	1
S(ER)	M^2/G	Esterase	U/G	Y = 46.657X	1
Esterase	U/G	S(ER)	M^2/G	Y = 0.0214X	1
S(ER)	M^2/G	NADPH-CCR	U/G	Y = 0.9829X	1
NADPH-CCR	U/G	S(ER)	M^2/G	Y = 1.0174X	1
S(OMIM)	M^2/G	MAO	U/G	Y = 0.6310X	1
MAO	U/G	S(OMIM)	M^2/G	Y = 1.5848X	1
S(IMIM)	M^2/G	S(IMIM)	U/G	Y = 6.8150X	1
CYOX	U/G	CYOX	M^2/G	Y = 0.1467X	1
S(PM)	M^2/G	5'NUC	U/G	Y = 26.267X	1
5'NUC	U/G	S(PM)	M^2/G	Y = 0.0381X	1

Monoamine oxidase (MAO) - a marker enzyme of the outer mitochondrial membrane (omim)

Figure 7.15 The worksheet converts units of marker enzyme activities to ER surface areas (m^2) and ER surface areas to units (U) of enzyme activities – for hepatocytes. The two prediction stacks evaluate the equations automatically (Original data adapted from Bolender et al., 1978, Losa et al., 1978, and Bolender et al., 1980; From Bolender, 2017).

Recall that biochemical homogeneity is being expressed as the:

Structure-Function Rule

$$\boldsymbol{f(x) = mx}\,. \qquad (7.8)$$

The equations used in the worksheets of Figure 7.15 are not playing the usual biological variation game. Instead, they are applying the rules under which biology allows variation to exist. We can write these equations because the ratios of the parts are highly conserved within and across species when they carry the same rules and instructions (genes). This reality can be used to our advantage by using equations to predict phenotypic expression upstream where we expect to encounter and connect to the rules being orchestrated by the genome.

7.3.1 Predicting Morphology from Morphology

In a control setting, for example, membrane organelles, occur in ratios specific to a given cell type. [Recall figures 3.7 and 3.8.]

Ratio Chain Rule (Morphology)

$$\boldsymbol{m_1 : m_2 : m_3, \ldots, m_n} \qquad (7.9)$$

When applied to organelle surface areas:

$$s_1 : s_2 : s_3, \ldots, s_n \ , \qquad (7.10)$$

the rule is expressed relative to the first item in the series:

$$\frac{s_1}{s_1} : \frac{s_2}{s_1} : \frac{s_3}{s_1} : , \ldots, \frac{s_n}{s_1} \ . \qquad (7.11)$$

We can simplify the notation of Ratio 7.11 as follows:

$$s'_1 : s'_2 : s'_3 : , \ldots, s'_n \ . \qquad (7.12)$$

By setting the first ratio in the series (chain) equal to 1 (s'_1/s'_1), a change made to this value can be applied to all the remaining items in the chain. Any member of the chain can become the first ratio.

The calculation worksheet generates the calculations automatically (Figure 7.16). Entering a seed value (into the first item in the chain ratio) equal to 4.6 m^2, would result in the following calculation:

$$4.6'_1 : 4.6 \times s'_2 : 4.6 \times s'_3 : , \ldots, 4.6 \times s'_n \ . \qquad (7.13)$$

Enter this value (4.6 m^2 of ER/gram of liver) into the space provided in the worksheet (Figure 7.16) and Excel calculates values for the remaining hepatocytic organelles listed in the membrane column. When the goal is diagnosis or prediction, these predicted values can be converted into mathematical markers and used as described in Chapter 4.

Figure 7.16 identifies unanswered questions of considerable interest. If a pattern exists across the morphology of the controls, how does the genome change this pattern when it's called upon to adapt? Why and how do parts change at different rates?

PREDICT SURFACE AREAS OF ORGANELLES IN THE HEPATOCYTE (RAT LIVER) - CORRECTED FOR SECTION RELATED

Generate Ratios				Enter x ↓		Enter x ↓		Enter x ↓	
	M^2/G	M^2/G	Ratio	er	4.6	ser	1	rer	1 go
Membrane	S(i)	S(j)	S(I)/s(j)						
er	13.800	8.650	1.595	1.000	4.600	2.604	2.604	1.622	1.622
ser	5.300	8.650	0.613	0.384	1.767	1.000	1.000	0.623	0.623
rer	8.510	8.650	0.984	0.617	2.837	1.606	1.606	1.000	1.000
go	0.556	8.650	0.064	0.040	0.185	0.105	0.105	0.065	0.065
pm	1.808	8.650	0.209	0.131	0.603	0.341	0.341	0.212	0.212
mim	11.680	8.650	1.350	0.846	3.893	2.204	2.204	1.373	1.373
omim	3.038	8.650	0.351	0.220	1.013	0.573	0.573	0.357	0.357
imim	8.650	8.650	1.000	0.627	2.883	1.632	1.632	1.016	1.016

TED BIASES - CORRECT NEW DATA BEFORE ENTERING

Enter x ↓		Enter x ↓		Enter x ↓		Enter x ↓		Enter x ↓	
go	1	pm	1	mim	1	omim	1	imim	1
24.820	24.820	7.633	7.633	1.182	1.182	4.542	4.542	1.595	1.595
9.532	9.532	2.931	2.931	0.454	0.454	1.745	1.745	0.613	0.613
15.306	15.306	4.707	4.707	0.729	0.729	2.801	2.801	0.984	0.984
1.000	1.000	0.308	0.308	0.048	0.048	0.183	0.183	0.064	0.064
3.252	3.252	1.000	1.000	0.155	0.155	0.595	0.595	0.209	0.209
21.007	21.007	6.460	6.460	1.000	1.000	3.845	3.845	1.350	1.350
5.464	5.464	1.680	1.680	0.260	0.260	1.000	1.000	0.351	0.351
15.558	15.558	4.784	4.784	0.741	0.741	2.847	2.847	1.000	1.000

Figure 7.16 In control cells, organelles can exist in well-defined proportions. Consequently, a panel of cell organelles can be predicted from a single estimate – e.g., when the surface of the er = 4.6 m^2, the ser = 1.767 m^2, et cetera (From Bolender, 2017).

Biology's structural design, which relies heavily on ratios, scales up and down the hierarchy of size. The liver, for example, contains four parenchymal cell types, wherein each cell displays organelles in proportions specific to its phenotype. In turn, intra- and intercellular proportions of organelles – along with the cell frequencies - define the phenotype of the cells making up the liver parenchyma. This pattern – defined by ratios – allows us to predict cellular events in the control liver parenchyma triggered by data coming from a single membrane surface area or marker enzyme activity (Figure 7.17). [Note that the membrane organelles being used in the figures were corrected for the section related biases.]

RAT LIVER - PREDICT RATIOS OF ORGANELLES - ONE CELL AT A TIME

		To calculate ratio			To calculate ratio	
	m^2/cm^3 paren	enter value ↓	%	m^2/cm^3 paren	enter value ↓	%
compartment i	**hepatocyte**	3.801	check	**endothelial cell**	1.000	check
pm	0.563	0.148	3.5%	0.116	0.116	22.0%
er	3.801	1.000	23.7%	0.134	0.134	25.5%
rer	2.412	0.635	15.0%	0.081	0.081	15.4%
ser	1.389	0.365	8.6%	0.053	0.053	10.1%
mi	3.825	1.006	23.8%	0.049	0.049	9.3%
omim	1.300	0.342	8.1%	0.020	0.020	3.8%
imim	2.525	0.664	15.7%	0.029	0.029	5.5%
go	0.170	0.045	1.1%	0.024	0.024	4.6%
ly	0.070	0.018	0.4%	0.012	0.012	2.3%
pinoves	0.011	0.003	0.1%	0.008	0.008	1.6%
Σ	16.066	4.227	100.0%	0.526	0.526	100.0%

	To calculate ratio			To calculate ratio	
m^2/cm^3 paren	enter value ↓	%	m^2/cm^3 paren	enter value ↓	%
kupffer cell	1.000	check	**fat-storing cell**	1.000	check
0.033	0.033	11.9%	0.054	0.054	28.9%
0.080	0.080	29.1%	0.045	0.045	23.9%
0.061	0.061	22.1%	0.036	0.036	19.1%
0.019	0.019	7.0%	0.009	0.009	4.9%
0.027	0.027	9.9%	0.020	0.020	10.8%
0.009	0.009	3.2%	0.007	0.007	3.9%
0.018	0.018	6.7%	0.013	0.013	6.7%
0.004	0.004	1.5%	0.003	0.003	1.3%
0.020	0.020	7.3%	0.001	0.001	0.5%
0.003	0.003	1.2%	0.000	0.000	0.1%
0.274	0.274	100.0%	0.187	0.187	100.0%

Figure 7.17 Predicting the surface areas of organelles in the cells populating the liver parenchyma (Original data adapted from Blouin eta al., 1977; From Bolender, 2017). Since biology arranges its cells and tissues by rule, we can use these rules - expressed as equations – to make predictions.

7.3.2 Predicting Biochemistry from Biochemistry

In a control setting, enzyme activities in hepatocytes display distinct relationships, which define ratios. Consequently, a string of ratios becomes predictive (Figure 7.18).

Ratio Chain Rule (Biochemistry)

$$\boldsymbol{b_1}:\boldsymbol{b_2}:\boldsymbol{b_3},\ldots,\boldsymbol{b_n} \tag{7.14}$$

When applied to units of enzyme activity:

$$u_1:u_2:u_3,\ldots,u_n\ , \tag{7.15}$$

the rule is expressed relative to the first item in the series:

$$\frac{u_1}{u_1}:\frac{u_2}{u_1}:\frac{u_3}{u_1}:,\ldots,\frac{u_n}{u_1} \tag{7.16}$$

Note that units of activity are related to a standardized mg of protein reference before assembling the ratios (Bolender, 2017). This is necessary because the mg protein reference behaves as a variable, being influenced by both biological and methodological variation. Standardization of enzyme assays may also be required (Figure 7.44).

Figure 7.18 uses published data (Amar-Costesec et al., 1974) to generate ratios that predict enzyme activities for 21 connected marker enzymes from a single control value. 5'nucleotidase, a marker enzyme for the plasma membrane of hepatocytes, serves as the example in the worksheet.

Notice how we can deal with the biological variability, which comes from the practice of relating enzyme activity to a mg of protein or gram of liver. Variability across animals and publications is minimized by using ratios of one activity to another instead of comparing concentrations. Variability coming from the data reference – the variable in the denominator – cancels out algebraically.

PURPOSE: PREDICT PANEL OF ENZYME ACTIVIES FROM A SINGLE ENZYME ASSAY — ASSUMPTION: ENZYMES ARE RELATED

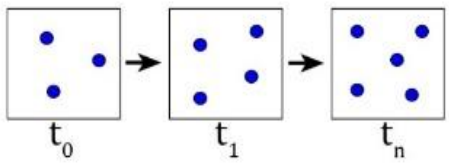
					Enter x ↓	
Enzyme	Abbreviation	Location	U/g Liver	ratio - in	5nuc	1 ac
5'-nucleotidase	5nuc	pm	11.300	100.000	1.000	1.000
acid phosphatase	acpase	lysosome	5.670	100.000	0.502	0.502
aldolase	ald	cytoplasmic	7.960	100.000	0.704	0.704
alkaline phosphatase	alkpase	pm	2.450	100.000	0.217	0.217
alkaline phosphodiesterase 1	alpd1	pm	17.500	100.000	1.549	1.549
aminopyrine demethylase	amdem	er	0.079	100.000	0.007	0.007
b-glucuronidase	bglur	lysosome	1.170	100.000	0.104	0.104
catalase	cata	peroxisome	47.600	100.000	4.212	4.212
cytochrome b5	cyb5	er	19.200	100.000	1.699	1.699
cytochrome oxidase	cyox	imim	18.900	100.000	1.673	1.673
cytochrome p 450	cyp450	er	21.700	100.000	1.920	1.920
esterase	est	er	257.000	100.000	22.743	22.743
fumarase	fum	mi	95.600	100.000	8.460	8.460
galactosyl transferase	galtrans	golgi	0.013	100.000	0.001	0.001
glucose-6-phosphatase	g6pase	er	20.200	100.000	1.788	1.788
glucuronyltransferase	glutrans	er	2.380	100.000	0.211	0.211
glutamine synthetase	glusyn	mi	8.610	100.000	0.762	0.762
monoamine oxidase	mao	omim	0.507	100.000	0.045	0.045
n-acetyl-b-glucosaminidase	nacebglu	lysosome	6.880	100.000	0.609	0.609
nadh cytochrome c reductase	nadhccr	imim	100.000	100.000	8.850	8.850
nadph cytochrome c reductase	nadphccr	er	3.980	100.000	0.352	0.352
nucleoside diphosphatase	nucdipase	golgi	100.000	100.000	8.850	8.850

Figure 7.18 The column at the right uses a control value (Enter x = 1) for a plasma membrane marker enzyme (5'nucleotidase) to predict the corresponding values for the enzymes listed at the left (Original data adapted from Amar-Costesec et al., 1974; From Bolender, 2017).

The point being made with these calculation worksheets (Figures 7.16 to 7.18) is that the structures and functions of cells and tissues follow exacting rules that produce distinct and reproducible patterns that collectively define phenotypes quantitatively. Recall that we first detected these patterns with organism codes (Figures 3.7 and 3.8). That observation provided a clue that led to the chain rules.

In experimental settings, this approach could be used to check for duplicate or missing genes, to monitor genetic regulation, to estimate rates of synthesis (rate constants), and to reconstruct metabolic tables for a host of phenotypic responses.

7.4 Level 3 – Organelle Changes



7.4.1 Detecting Changes – Separately and Together

In the sciences, theory structure becomes the defining driver of discovery. Under reductionist theory, we take biology apart and then characterize the remaining parts morphologically and biochemically. When we plot changes in morphology versus biochemistry, however, the results can be disappointing – the $R^2 = 1$ or $R^2 \approx 1$ equations fail to appear. If, instead, we put biology back together within the framework of complexity theory, we get the $R^2 = 1$ and $R^2 \approx 1$ equations. Although we take biology apart physically, we put it back together mathematically – according to biology's rules.

7.4.1.1 Morphology

Consider liver hepatocytes. Stereological estimates of ER membrane surface areas, for example, detect changes (increases and decreases) in response to phenobarbital treatment. Figure 7.19 indicates that following the initial increase, the ER shows little change and even suggests a slight decrease when related to a gram of liver. Although the hepatocytes are filling up with ER membranes designed to metabolize the drug, we can't detect this increase because each day we are slowing sliding - deeper and deeper - into the concentration trap wherein fewer and fewer hepatocytes are needed to fill up a gram (or cm^3) of liver. Fewer cells means less membrane surface area per gram of liver. Figure 7.19 illustrates the regrettable outcome of this black box approach to detecting a biological change.

S/g liver			
days	er/g	ser/g	rer/g
0	8.65	2.96	5.80
0.67	11.61	4.68	6.94
2	11.97	5.20	6.40
5	11.14	5.83	4.73

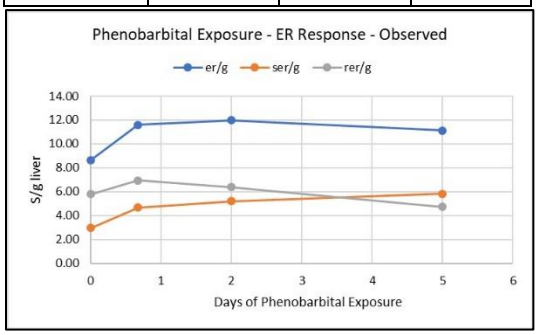


Figure 7.19 ER surface areas of hepatocytes estimated stereologically and related to a gram of liver (Original data adapted from Stäubli et al., 1969); From Bolender, 2018).

enz/g liver			
days	cytoc p450/g	n-demeth/g	nadph-ccr/g
0	9.2825	1.7443	0.0990
0.67	13.9900	6.0097	0.0992
2	21.0447	8.0224	0.1154
5	33.7658	12.7447	0.1772

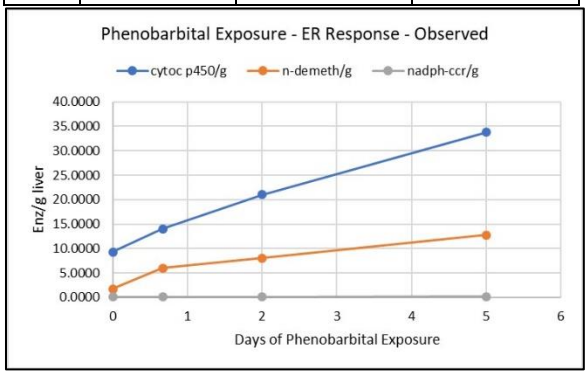


Figure 7.20 Enzyme activities of hepatocytes estimated biochemically (Original data adapted from Stäubli et al., 1969); From Bolender, 2018).

7.4.1.2 Biochemistry

Notice in Figure 7.20 that the biochemical measures of two out of three membrane bound marker enzymes clearly increase in response to phenobarbital exposure. However, the slopes of the enzyme activity curves are diminished because of the diluting effect of the diminishing numbers of hepatocytes contained within a gram of liver. Both the morphological and biochemical data inevitably fall into the same concentration trap, but it's harder to see biochemically because the enzyme activities are increasing faster than the ER membranes and faster than the loss of activity attributed to hepatocytes exiting the gram of liver.

Trying to do biochemistry without morphology or vice versa, comes with the risk of having to connect single, isolated pieces of information to the constantly shifting complexity of a biological change.

7.4.1.3 Morphology vs. Biochemistry

If we combine the data of Figures 7.19 and 7.20 to look for relationships of structure to function (i.e., plot ER surface areas against marker enzyme activities), we get disappointing results with R^2s = 0.48 and 0.54 (Figure 7.21). Now our analysis suggests that both enzymes (cytochrome P450 and n-demethylase) also appear to be decreasing in response to phenobarbital. Such results tell us that this is not the way to find a rule that defines the relationship of structure to function in cells.

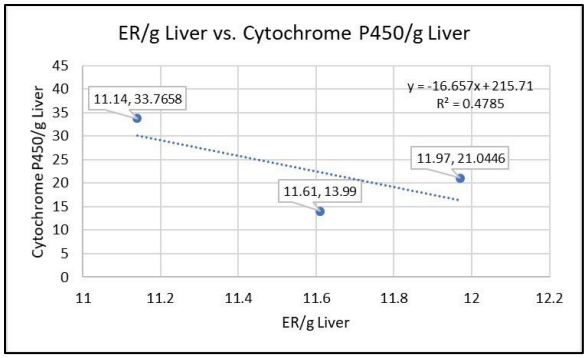


Figure 7.21 ER surface areas plotted against marker enzyme activities fail to detect a crisp quantitative relationship (Original data adapted from Stäubli et al., 1969).

By operating under the constraints of a weak theory structure (reductionism), we are left with the task of dealing with the consequences of the results.

7.4.1.4 Structure-Function Rules

If, however, we relate each experimental time point listed in Figure 7.20 to 1 m^2 of ER surface area by calculating enzyme densities, then the resulting points fit linear regressions, deliver the expected R^2s, and we can see the expression of biology's structure-function change rule – one for each membrane-enzyme combination (Figure 7.22).

By switching to complexity, we can discover how hepatocytes respond to phenobarbital by changing the relationship of structure to function in their ER membranes. Since biology manages its structural and functional responses to the drug holistically, it increases both the concentration of its marker enzymes per unit of ER surface area (shown in Figure 7.22) and the amount of ER membrane in its hepatocytes (shown in Figure 5.4 and 7.31).

Here we see a biological complexity expressed as an adaptive response operating locally to increase the concentration of a membrane-bound enzyme and globally as an increase in the total amount of hepatocytic ER membrane in its parent organ - the liver.

In this experiment, biology used two mechanisms (sets of rules) to increase the capacity of its drug metabolizing enzymes. However, this represents only one part of the much larger story of change. Biology also has rules for determining the rate at which the induced (and non-induced) enzymes are added to the ER membranes and for setting the turnover rates of the preexisting control membranes (Bolender and Weibel, 1973).

	ED	ED	ED
days	nadph-ccr/S(er)	cytoc p450/S(er)	n-demeth/S(er)
0	0.0114	1.0731	0.2017
0.67	0.0085	1.2054	0.5178
2	0.0096	1.7580	0.6702
5	0.0159	3.0319	1.1444

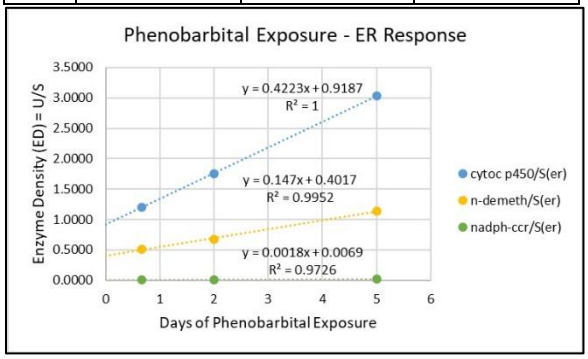


Figure 7.22 Local modifications to membranes represent complex events, which biology defines as distinct relationships of structure to function (Original data adapted from Stäubli et al., 1969); From Bolender, 2018). Notice that biology defines its relationship of structure to function with ratios (EDs).

7.4.1.5 States: Steady (Co) → Transitional → Steady (Ex)

Recall that the change model introduced in Figure 1.1 included two steady states (control and experimental) separated by a transitional state (change). Figure 7.22, which characterizes a transitional state, displays persistent increases in three enzyme densities with distinctly different slopes (rate constants) in response to the drug phenobarbital. A new steady state (Ex) appears to be nowhere in sight. However, ratios characterizing the enzyme densities at days 0 (Co) and PB days 2, 5 (Ex) are distinctly different (Figure 7.23). That's a clue.

7.4.2 Enzyme Density Ratios

In biology, quantitative patterns – expressed as ratios - seem to exist everywhere. For example, the enzyme density ratios of two different marker enzymes induced by phenobarbital change momentarily and then establish a new steady state relationship (Figure 7.23).

days	ED 1-2	ED 1-3	ED 2-3
0	5.321622	0.937619	0.17619
0.67	2.327928	141.007	60.57189
2	2.62322	182.2991	69.49441
5	2.649411	190.5576	71.92454

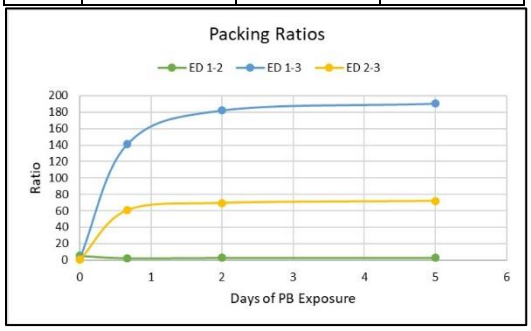


Figure 7.23 The relative amounts (ratios) of cytochrome P450, n-demethylase, and NADPH cytochrome c reductase change in response to phenobarbital. ED 1-2: n-demethylase vs cytochrome P450; ED 1-3: NADPHCCR vs cytochrome P450; ED 2-3: NADPHCCR vs n-demethylase (Original data adapted from Stäubli et al., 1969); From Bolender, 2018).

This tells us that hepatocytes respond to phenobarbital by reprogramming the relationship of one ER marker enzyme to another. In other words, the membrane recipe changes. This means that a change in the ER involves two complex modifications – adaptation (new ED packing ratios) and production (more ER).

7.4.3 Transitional State

The transitional state represents a time during which major changes occur. It's a seemingly chaotic period when many variables change at different rates across multiple levels of complexity.

During the transition, the control recipe for the ER membranes is being gradually replaced by a new one designed to eliminate the phenobarbital. What's the new recipe and how do we detect and characterize it?

7.4.3.1 Looking for the New Recipe

In response to phenobarbital exposure, hepatocytes change the steady state of its ER enzyme densities from control to experimental (PB), as shown in (Figure 7.24). After five days of exposure to the drug, the cells roughly doubled, for example, the relative amount of n-demethylase activity associated with a square meter of ER membrane (Figure 7.24).

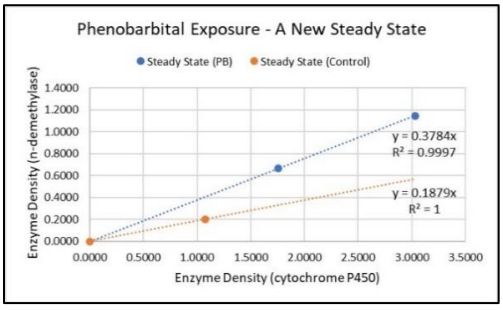


Figure 7.24 The two equations identify steady states before (control) and after exposure to phenobarbital (PB). The PB

curve uses data taken from Figure 7.22 (Original data adapted from Stäubli et al., 1969).

Since the enzyme density curves for n-demethylase and cytochrome P450 are increasing at different rates (Figure 7.22), how do we show that a new steady state (PB) exists for these two enzymes between PB days 2 and 5? In other words, how does biology embed a steady state (Figure 7.23) into a transitional state (Figure 7.22)? Is this a hard question? No.

We need do little more than watch biology doing its job. We know that an enzyme density adheres to the structure-function rule ($f(x) = mx$) at each experimental time point. By plotting one enzyme density against another (Figure 7.25), we can expect to get change equations ($f(x) = mx + b$) – (top and middle panels) – until the new recipe is established ($f(x) = mx$) – (bottom panel). The change equations (top and middle panels) miss the origin, whereas the bottom one (day 2 to day 5) passes through it. This tells us that at days 2 to 5 the enzyme densities of n-demethylase and cytochrome P450 are no longer changing relative to each other and have established a new steady state (bottom). This signals yet another interpretation of our general structure-function rule:

Recipe Detection Rule:

$$\boldsymbol{f(x) = mx}\text{, where } \boldsymbol{m = ED_i/ED_j}\text{,} \quad (7.17)$$

and i and j identify two different enzyme densities (ED). For phenobarbital exposure (PB Day 2 and 5):

$$y = 0.3784x \quad (7.18)$$

and the line passes through origin (Figure 7.25). This result indicates that a new and stable relationship exists between n-demethylase and cytochrome P450, wherein the relative amount of n-demethylase increased 100% compared to the control (Figure 7.24).

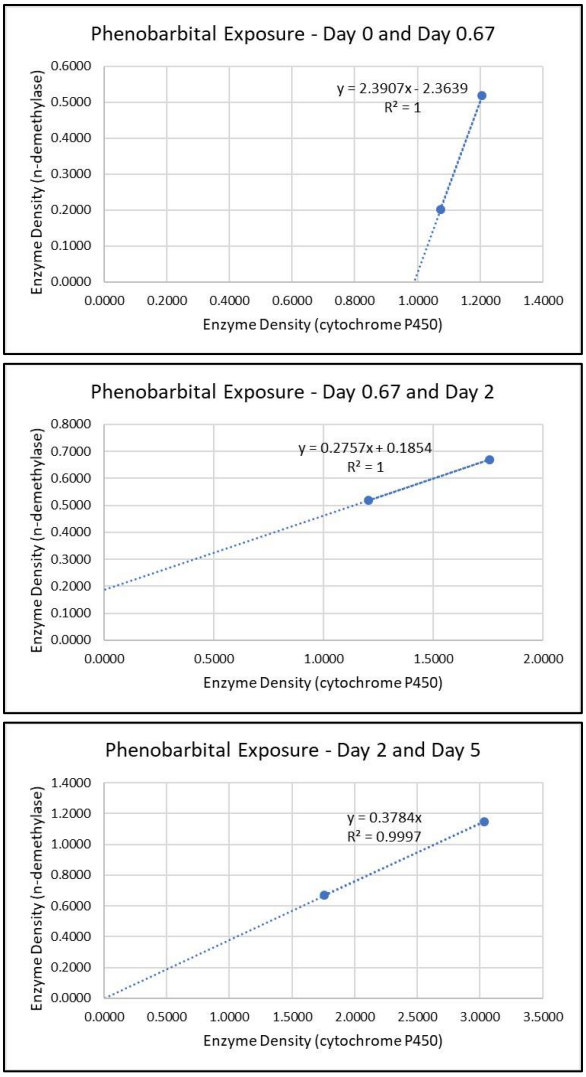


Figure 7.25 Biology adapts to phenobarbital exposure by changing hepatocytic relationships of structure to function according to its rule-based recipe for metabolizing the drug (Original data adapted from Stäubli et al., 1969).

Within two days, hepatocytes were able to recognize the phenobarbital threat, figure out how to respond, and apply a solution that included making changes to their ER membranes - both in their molecular composition and amount. Figures 7.22 and 7.25 identify the transitional stage of a membrane change as a two-part process: dynamic (transition) and dynamic equilibrium (steady state).

The complexity of this membrane change encourages speculation. Certain disorders

might be explained by an inability of cells to establish a steady state wherein they remain trapped in the limbo of a transitional zone. Such a scenario might help to explain the large numbers of abnormal mathematical markers associated with the disease process (see, for example, Figures 3.25 and 4.14).

7.4.3.3 Finding the Recipe Rule

The enzyme densities define a membrane recipe as a ratio of multiple marker enzyme activities. Moreover, these ratios, which define recipe rules, can be translated into mathematical markers and used to detect specific responses to specific exposures.

Recipe Rule

$$\boldsymbol{a(ED_1):b(ED_2):c(ED_3)\ldots x(ED_n)} \quad (7.19)$$

For the phenobarbital study, we can use the EDs of the control and experimental steady states (Figure 7.22) to identify the change in the recipe:

Steady State (Co)

$$demER(ED_1):cy450ER(ED_2):nadphccrER(ED_3)(7.20)$$
$$\mathbf{demER0.2:cy450ER1.1:nadphccrER0.011}$$

Steady State (Ex)

$$demER(ED_1):cy450ER(ED_2):nadphccrER(ED_3)(7.21)$$
$$\mathbf{(demER1.1:cy450ER3.0:nadphccrER0.016}$$

When expressed as a mathematical marker, the recipe rule speeds the task of finding similar patterns in large data sets (Figure 7.26). Notice that the mathematical markers at PB days 2 to 5 detect the new membrane recipe for phenobarbital – as predicted by Figures 7.23 and 7.25.

Enzyme Densities to Mathematical Markers			
PB Days	ED dem	ED cytocp450	ED nadphccr
0	dem1cytocp4505.3nadphccr0.057		
0.67	dem1cytocp4502.3nadphccr0.017		
2	**dem1cytocp4502.6nadphccr0.014**		
3	**dem1cytocp4502.6nadphccr0.014**		
4	**dem1cytocp4502.6nadphccr0.014**		
5	**dem1cytocp4502.6nadphccr0.014**		

Figure 7.26 Enzyme densities can be translated into mathematical markers. This new category of mathematical markers – based on enzyme densities – might be especially helpful as we begin to decode the orchestration of the DNA networks. Note that the markers – for purpose of illustration - were not converted to decimal repertoire values (Original data adapted from Stäubli et al., 1969).

By converting the results of the recipe rule into mathematical markers, we add a new data type that improves our chances of finding diagnostic, predictive, and interpretable patterns. This becomes important when trying to generalize phenotypic responses to a wide range of exposures. Moreover, when working out the rules operating during gene expression in the phenotype timing is everything. From the way the recipe rules operate (Figure 7.25 and 7.26), we can now approach the transitional state as a series of events wherein different things happen at different times. In other words, we're learning how to play the complexity game.

7.4.4 The Black Box

Let's pause for a moment and pull together what we are learning about how biology changes. We can view a biological change from two perspectives – one copied from physics and chemistry and the other copied from physics, chemistry, and biology. Both approaches begin by taking biology apart. We have two choices. We can learn everything we want to know by studying the parts or we can use the parts to

put biology back together and learn whatever biology is willing to teach us.

What, for example, did we learn about change from the parts before and after putting biology back together?

7.4.4.1 Simple Change (Box Closed)

The phenobarbital study (Stäubli et al., 1969), which took the parts approach, reported the following changes to the ER membranes of hepatocytes.

End State (PB Day 5)

Compared to the control, the specific surface area of the ER increased at day 0.67 ($P<0.01$) and at days 2 and 5 ($P<0.001$); the RER at day 2 ($P<0.01$) and the SER at day 0.67 ($P<0.01$) and days 2 and 5 ($P<0.001$). Moreover, cytochrome P450 and n-demethylase were significantly different ($P<0.001$) from the control at all the experimental time points.

7.4.4.2 Complex Change (Box Open)

During the transitional state when structural and functional changes occur, several biological rules were in play simultaneously.

Transitional State (PB Days 2 to 5):

- Each new enzyme density expresses the **structure-function rule**.
- Each enzyme density increased over time, adhered to the **structure-function change rule**, and displayed a different rate constant (slope).
- The enzyme densities indicated that the concentrations of the marker enzymes in the ER membranes continued to increase through the experiment.
- The new **recipe rule**, which defines the relative packing of the enzymes in the ER membrane, detected a new membrane recipe at PB day 2.
- When expressed as mathematical markers, the enzyme densities detected the new **recipe** at PB days 2 to 5.
- Beginning at PB day 2, the hepatocytes added 5.86 m^2 of new ER to the liver per day.

Notice that simplification told us the ending of the phenobarbital story, whereas complexity told us how the ending came about.

7.4.5 Moving Upstream

7.4.5.1 Predicting mRNA from Enzyme Densities

Once we identify the players and the sequence of a change over time, we can begin to use such phenotypic information to look for and predict the controlling events occurring upstream at the levels of RNA and DNA (Bolender, 2018).

Starting with data published in the Stäubli paper (1969), we can generate enzyme densities and use them to predict the appearance of mRNA in liver hepatocytes in response to the phenobarbital induction (Figure 7.27; see also Bolender, 2018). Although these results are preliminary, they serve to illustrate a potential strategy for linking *in situ* enzymes to their parental mRNAs.

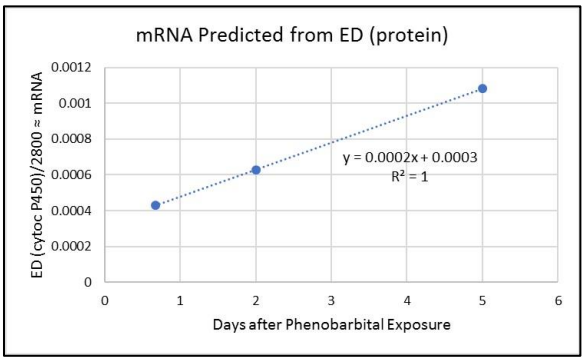
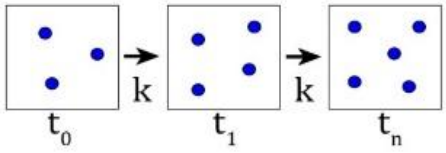


Figure 7.27 This predicted curve for mRNA runs roughly parallel to the original cytochrome P450 curve shown in Figure 7.22 (Original data adapted from Stäubli et al., 1969); From Bolender, 2018). Once again, we find evidence for a rule-based change, this time extended to mRNA.

7.5 Level 4 – Rates of Change



7.5.1 Rate Constants

Biology defines changes mathematically as a complex relationship of structure to function, which we can detect over time with rate constant equations displaying $R^2 = 1$ or $R^2 \approx 1$.

To illustrate the calculation, we will use rate equations to predict the missing EDs at day 3 for three marker enzymes (Figure 7.28).

7.5.1.1 Integrated Rate Law

We begin with the rate law:

$$[A] = [A_0] \pm kt \text{ (for (+) and (-) slopes)} \quad (7.22)$$

Wherein the linear plot is $[A]$ $versus$ t .

Substituting enzyme densities (ED) and a positive slope gives:

$$[ED_3] = [ED_0] + kt, \quad (7.23)$$

where:

$[ED_i] = \frac{Units}{Surface}$ *at time* i $=$ *concentration* $[A]$,

k *= slope = rate constant (units = concentration/time),*

t $=$ *time*.

7.5.1.2 Prediction - Day 3 of PB induction

Note that day 3 becomes Day 2 here because Day 1 = Day 0) in the rate equation (Original data adapted from Stäubli et al., 1969; From Bolender, 2018).

Cytochrome P450

$$[ED_3] = [ED_0] + kt \quad (7.24)$$

$$[\ \langle\!\langle ED \rangle\!\rangle\_3\ ] = [1.341] + [0.4223/d] \times 2\ d$$

$$[ED_3] = 2.1856$$

n-demethylase

$$[ED_3] = [ED_0] + kt \quad (7.25)$$

$$[\ \langle\!\langle ED \rangle\!\rangle\_3\ ] = [0.5487] + [0.147/d] \times 2\ d$$

$$[ED_3] = 0.8427$$

NADPH Cytochrome c reductase

$$[ED_3] = [ED_0] + kt \quad (7.26)$$

$$[\ \langle\!\langle ED \rangle\!\rangle\_3\ ] = [0.0087] + [0.0018/d] \times 2\ d$$

$$[ED_3] = 0.0123$$

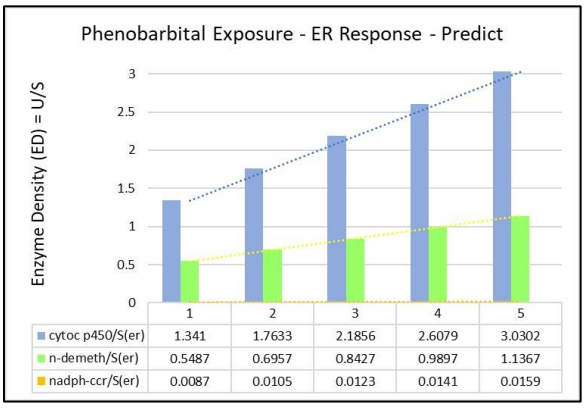
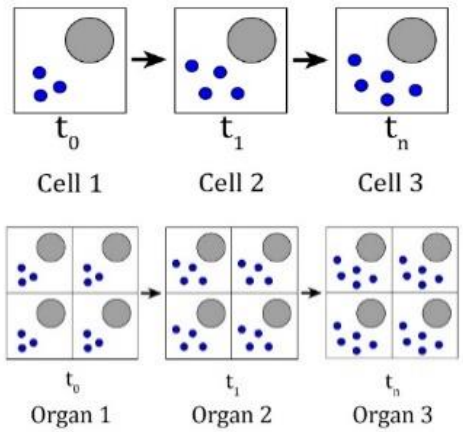


Figure 7.28 The rate equations predicted the missing concentrations (EDs) at days 3 and 4 (Original data adapted from Stäubli et al., 1969); From Bolender, 2018).

By tying the calculation of rate constants to a common reference (1 m^2), we can use the ratio change rules to predict combinations of enzymes, membranes, and organelles to quantify the metabolic activities of a cell *in situ* (recall Figures 7.15 to 7.18). Such a strategy

may encourage attempts to simulate portions of the metabolic chart in living cells.

7.6 Levels 5, 6 – Cell and Organ Changes



7.6.1 Detecting Biochemical Changes – Per Gram, Per Cell, and Per Organ

When left uncorrected, the concentration data (expressed per gram of liver) underestimates both the rate (slope) and amount (activity) of the cellular changes induced by phenobarbital. Moreover, these errors are not trivial.

A biological change becomes a lesson in complexity. Typically, it involves organelle and enzyme concentrations, cell frequencies, cell shapes, cell volumes, organ volumes, et cetera. We can, for example, use the CCC equation (2.28) to show how such variables combine to produce an experimental outcome (Figure 7.29).

Results expressed as absolute values (per liver – without cell loss/gram liver) detect a greater change compared to those based on concentrations (with cell loss/gram of liver). At day 5 of phenobarbital treatment, for example, the corrected value for cytochrome P450 (per liver) is 60% larger than that of the original (per gram of liver). Larger differences, of course, tend to result in better P values.

	Original	Corrected (CCC)
	Cell loss /g Liver	No cell loss/g Liver
PB days	cytocp450/g Liver	cytocp450/Liver
0	9.283	9.283
0.67	13.990	21.085
2	21.045	31.657
5	33.766	54.177

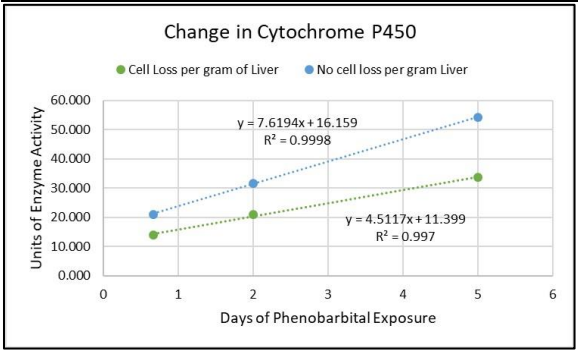
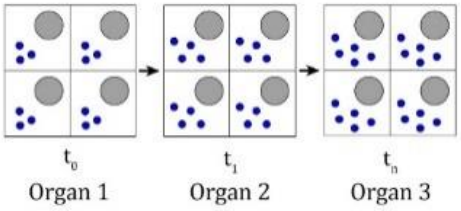


Figure 7.29 Notice how concentration data diminishes the slope of the curve (green line) and underestimates the extent of the change (blue line) by about 60%. Original data adapted from Stäubli et al., (1969); (From Bolender, 2018).

7.7 Level 6 – Organ Changes



7.7.1 Same Data - Different Results

A biological change puts a host of interacting variables in play, including molecules, organelles, cells, and organs. Cell volume, for example, is one of the most important. Figure 7.30 shows what happens when we ignore changes in cell volume (per gram of liver) or account for them (per liver). Notice that the two estimates differ by about 60%. This

example and the one in Figure 7.29 help to explain why so much disagreement exists between similar studies in the biology literature.

PB Exposure				
Days	Liver (g)	V Hep/g	V Hep/100gbw	V Hep/Liver
0	1.00	1.00	1.00	1.00
0.67	1.33	1.05	1.08	1.39
2	1.20	1.09	1.24	1.31
5	1.54	1.08	1.38	1.67

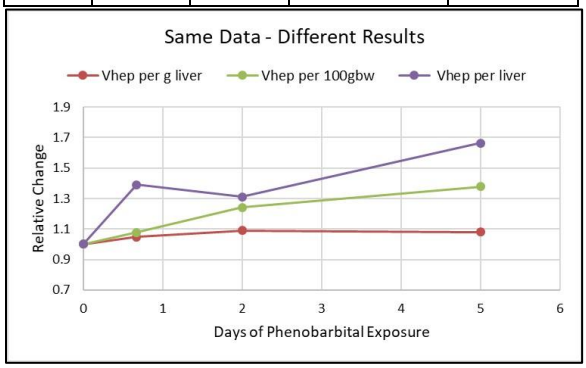


Figure 7.30 Detecting a biological change depends on the number of active variables accounted for in the data analysis (Original data adapted from Stäubli et al., 1969; From Bolender, 2018). By including more of the variables in play, we get a better result. Notice that each data type generates a distinctly different interpretation of the same experiment.

7.7.2 Changes in Organelles Expressed per Organ

Given the limitations imposed on hierarchy equations by the presence of multiple volume distortions (Equation 5.3), we currently lack a dependable method for estimating a key piece of information - the total (absolute) surface area of a membrane compartment in an organ. Recall that such information can also tell us what's happening in an average cell – when the number of cells in the population remains constant.

When morphology can't solve a problem alone or gets into trouble, we now have the option of turning to biochemistry for help. Since we know how biology instigates a change across multiple levels of complexity by rule, we can solve our absolute data problem by applying one of our new rule-based equations.

If we plug the absolute value for an ER marker enzyme activity (U_{total}) and its membrane density (MD) into Equation 7.27, we can solve for the absolute amount of ER membrane surface area in an organ. Figure 7.31 includes the data, equation, and calculations (for PB day 5).

$$S_{total} = MD \times U_{total} \qquad (7.27)$$

$$S_{er,liver} = MD_{S(er)/cytoP450} \times U_{cytoP450/liver}$$

$$S_{er,liver} = 0.3298 \times 330.3717 = 108.9651 \text{ m}^2$$

Data			
Phenobarbital Exposure	cytoP450	Ser/ cytocP450	m^2 (Total)
Days	U/Liver	MD	S(er)
0	58.8810	0.9319	54.8683
0.67	117.6576	0.8296	97.6115
2	160.6590	0.5688	91.3890
5	330.3717	0.3298	108.9651

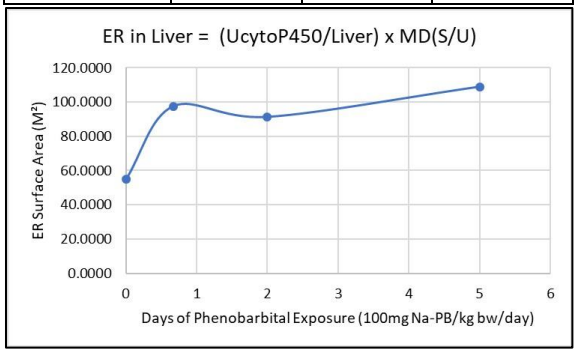


Figure 7.31 Biochemical data coupled with membrane densities (MD) can deliver estimates for the total ER surface area of hepatocytes in the liver with a minimum amount of stereologically related bias. This becomes a promising candidate to replace the troublesome hierarchy equation (Original data adapted from Stäubli et al., 1969). Equation 7.27 suggests a rule-based approach to two long-standing limitations of biological stereology - the volume distortions of hierarchy equations and the assumptions surrounding estimates of average cell data. Note: Similar calculations based on the other two marker enzymes give identical values for total S(er).

The results shown in the figure indicate that the amount of hepatocytic ER in the control liver doubles in response to the phenobarbital

exposure in about 16 hours, but the changes in ER membranes do not parallel those of its marker enzyme activities (Figure 7.20). Why?

Given the linear order detected at the earlier levels of complexity with $R^2 \approx 1$ equations (Figures 7.15, 7.22, 7.29), why do we end up with a nonlinear curve for the total ER surface area in the liver (Figure 7.31)? Such an irregular curve tells us that the ER membranes are undergoing multiple changes at the same time. While new membranes are being added (the phenobarbital induced ones), the original ones (control) are most likely being recycled by autophagic vacuoles and dense bodies.

This pattern of membrane removal becomes more pronounced after secession of drug treatment (Bolender and Weibel, 1973). In both studies, a dip in the membrane surface area occurs at day 2, which for the recovery paper corresponds to a spike in the amount of autophagic vacuoles and dense bodies in the hepatocyte cytoplasm.

Figure 7.31 indicates that the total changes in ER surface area follows a varied course. However, the recipe detection rule (Figure 7.25) indicates that linearity is reestablished at days 2 and 5 of phenobarbital exposure. Using the data in Figure 7.31, we can predict values for the intervening PB days 3 and 4. This allows us to characterize the increase in total ER surface area in the liver with a $R^2 \approx 1$ equation (Figure 7.32) where y equals the ER Surface area and x the days of phenobarbital exposure:

$$y = 5.86x + 79.67\ . \qquad (7.28)$$

Figure 7.32 shows that between days 2 and 5 the total hepatocytic surface area increases by 5.86 m^2 per day. Using the same recipe, hepatocytes make a new – but highly specialized - batch of ER membranes each day. In effect, the membrane becomes the hepatocyte's solution to the phenobarbital problem.

PB Exposure					
Day	Surface per liver	S ER	Step change	Step change	New ER per day
2	91.39	100%	97.25	100%	5.86 m^2
3	97.25	106%	103.11	106%	5.86 m^2
4	103.11	113%	108.97	106%	5.86 m^2
5	108.97	119%	114.82	106%	5.86 m^2

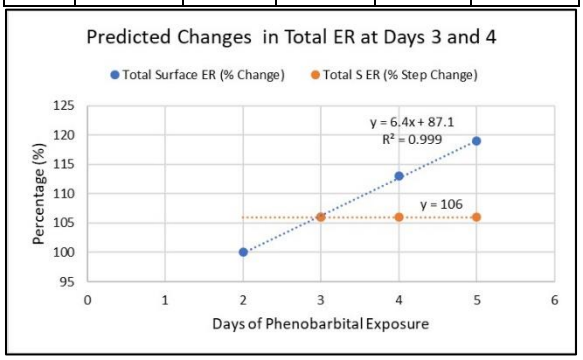


Figure 7.32 In response to phenobarbital exposure, hepatocytes change the composition of their ER membranes and the rate at which the new PB metabolizing membranes are produced (5.86 m^2/day).

7.7.3 Same Data - Same Results

In a rule-based system, the same data interpreted with different methods should give the same result. We can test this assumption by comparing the results of the CCC (Equation 2.23) and MD (Equation 7.27) methods, using the phenobarbital study to detect changes in the total liver ER. Figure 7.33 shows that both methods produce the same result.

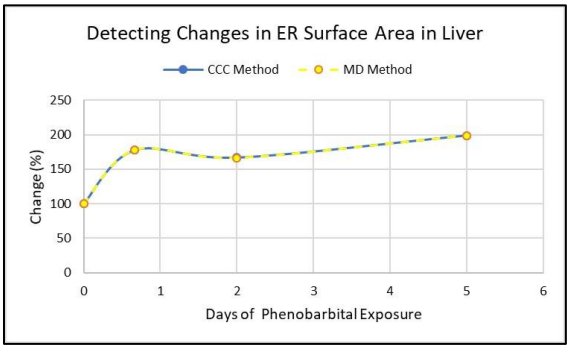


Figure 7.33 When related to the liver, two rule-based methods (CCC and MD) detect the same phenobarbital-induced changes in hepatocytic ER (Original data adapted from Stäubli et al., 1969).

7.8 Two Theory Structures Compared

The theory structure provides the operating system under which we explore biology. The primer argues that we need two, one to take biology apart (reductionist theory) and a second one to put biology back together (complexity theory). Figure 7.34 illustrates the types of outcomes these theory structures can deliver. When used together, they allow us to solve the phenotype.

7.8.1 Descriptive Summary of the Phenobarbital Study

REDUCTIONISM	**COMPLEXITY**
Subject: Detect significant changes in ER membranes	Subject: Explain changes in ER membranes
Strategy: Take biology apart, characterize parts	Strategy: Copy and apply biology's rules
Proficiency: Statistics	Proficiency: Biology's Package of Skills
Analysis: Black Box Model	Analysis: Biology Model
When related to specific dimensions (per 100 gbw):	When subject to biology's rules, the following apply:
1. The ER membranes were significantly different from the control at: 1.1. PB Day 0 38.41 m^2/100gbw 1.2. PB Day 0.67: P < 0.01 52.97 m^2/100gbw 1.3. PB Day 2: P < 0.001 60.54 m^2/100gbw 1.4. PB Day 5: P < 0.001 63.19 m^2/100gbw 2. NADPH-cytochrome c reductase activities significantly different from the control at: 2.1. PB Day 0 0.2100 units/100gbw 2.2. PB Day 0.67: P >0.05 0.2163 units/100gbw 2.3. PB Day 2: P < 0.001 0.2789 units/100gbw 2.4. PB Day 5: P < 0.001 0.4803 units/100gbw 3. N-demethylase activities were significantly different from the control at: 3.1. PB Day 0 3.7000 units/100gbw 3.2. PB Day 0.67: P < 0.001 13.1017 units/100gbw 3.3. PB Day 2: P < 0.001 19.3806 units/100gbw 3.4. PB Day 5: P < 0.001 34.5432 units/100gbw 4. Cytochrome P450 activities were significantly different from the control at: 4.1. PB Day 0 19.6900 units/100gbw 4.2. PB Day 0.67: P < 0.001 30.4998 units/100gbw 4.3. PB Day 2: P < 0.001 50.8396 units/100gbw 4.4. PB Day 5: P < 0.001 91.5191 units/100gbw	1. Level 1 Complexity **1.1. Mathematical markers** and **connection ratios** are shared across: 1.1.1. Animals 1.1.2. Cells of the liver parenchyma 1.1.3. Publications **1.2. Ratio Chain Rule** (Morphology) 1.2.1. Cell components occur in ratios specific to a given cell type: $\boldsymbol{m_1}:\boldsymbol{m_2}:\boldsymbol{m_3},\ldots,\boldsymbol{m_n}$ **1.3. Ratio Chain Rule** (Biochemistry) 1.3.1. Chemical constituents occur in ratios specific to a given organelle, compartment, or cell type: $\boldsymbol{b_1}:\boldsymbol{b_2}:\boldsymbol{b_3},\ldots,\boldsymbol{b_n}$ 2. Level 2 Complexity **2.1. Structure-Function Rule** 2.1.1. Each data point defines a relationship of structure to function expressed as enzyme or membrane density: $\boldsymbol{f(x)=mx}$ 2.1.2. Enzyme Density: **ED = U/S** Co Day 0 n-demethylase-ER ED = 0.2017 PB Day 0.67 n-demethylase-ER ED = 0.5178 PB Day 2 n-demethylase-ER ED = 0.6702 PB Day 5 n-demethylase-ER ED = 1.1444 2.1.3. Enzyme Density: **ED = U/S** Co Day 0 cytochrome P450-ER ED = 0.2017 PB Day 0.67 cytochrome P450-ER ED = 0.5178 PB Day 2 cytochrome P450-ER ED = 0.6702 PB Day 5 cytochrome P450-ER ED = 1.1444 2.1.4. Enzyme Density: **ED = U/S** Co Day 0 NADPHCCR-ER ED = 0.0114

PB Day 0.67 NADPHCCR-ER
ED = 0.0085
PB Day 2 NADPHCCR-ER
ED = 0.0096
PB Day5 NADPHCCR-ER
ED = 0.159

2.2. Biochemical Homogeneity Test

$[\boldsymbol{S_i} \times \boldsymbol{ED_i}] + [\boldsymbol{S_j} \times \boldsymbol{ED_j}] = \boldsymbol{U_{total}}$

$[1.90 \times ED_{ser}] + [2.97 \times ED_{rer}] = 27.421$

$[1.44 \times ED_{ser}] + [2.88 \times ED_{rer}] = 24.267$

$ED_{rer} = 5.54$

$ED_{ser} = 5.77$

2.3. Recipe Detection Rule:

$\boldsymbol{f(x) = mx}$, where $\boldsymbol{m} = \Delta \boldsymbol{ED_1}/\Delta \boldsymbol{ED_2}$

For PB Day 2 and 5

$y = 0.3784$ (passes through origin)

2.4. Recipe Rule:

$\boldsymbol{a(ED_1)}: \boldsymbol{b(ED_2)}: \boldsymbol{c(ED_3)} \ldots \boldsymbol{x(ED_n)}$

Control (Mathematical Marker)

$DemER(ED_1): cy450ER(ED_2): nadphccrER(ED_3)$

$DemER0.2cy450ER1.1nadphccrER0.01$

Experimental (Mathematical Marker)

$DemER(ED_1): cy450ER(ED_2): nadphccrER(ED_3)$

$DemER1.1cy450ER3.0nadphccrER0.02$

3. Level 3 Complexity

3.1. Structure-Function Change Rules:

$\boldsymbol{f(x) = mx + b}$

3.1.1. N-demethylase-ER

$y = 0.147x + 0.4017$

3.1.2. Cytochrome P450-ER

$y = 0.4223x + 0.9187$

3.1.3. NADPHCCR-ER

$y = 0.0018x + 0.0069$

4. Level 4 Complexity

4.1. Rate Equation Rule:

$[\boldsymbol{A}] = [\boldsymbol{A}]_{\boldsymbol{0}} \pm \boldsymbol{kt}$

4.1.1. N-demethylase-ER (PB Day 0.67)

$[ED_{0.67}] = [ED_0] + kt$

$[ED_{0.67}] = 0.5178$

4.1.2. Cytochrome P450-ER (PB Day 0.67)

$[ED_{0.67}] = [ED_0] + kt$

$[ED_{0.67}] = 01.2054$

4.1.3. NADPHCCR-ER (PB Day 0.67)

$[ED_{0.67}] = [ED_0] + kt$

$[ED_{0.67}] = 0.0085$

4.1.4. N-demethylase-ER (PB Day 2)

$[ED_2] = [ED_0] + kt$

$[ED_2] = 0.6702$

4.1.5. Cytochrome P450-ER (PB Day 2)

$[ED_2] = [ED_0] + kt$

$[ED_2] = 1.7580$

4.1.6. NADPHCCR-ER (PB Day 2)

$[ED_2] = [ED_0] + kt$

$[ED_2] = 0.0096$

4.1.7. N-demethylase-ER (PB Day 3)

$[ED_3] = [ED_0] + kt$

$[ED_3] = 0.8427$

4.1.8. Cytochrome P450-ER (PB Day 3)

$[ED_3] = [ED_0] + kt$

	$[ED_3] = 02.1856$ 4.1.9. NADPHCCR-ER (PB Day 3) $[ED_3] = [ED_0] + kt$ $[ED_3] = 0.0123$ 4.1.10.N-demethylase-ER (PB Day 4) $[ED_4] = [ED_0] + kt$ $[ED_4] = 0.9897$ 4.1.11.Cytochrome P450-ER (PB Day 4) $[ED_4] = [ED_0] + kt$ $[ED_4] = 02.6079$ 4.1.12.NADPHCCR-ER (PB Day 4) $[ED_4] = [ED_0] + kt$ $[ED_4] = 0.0141$ 4.1.13.N-demethylase-ER (PB Day 5) $[ED_5] = [ED_0] + kt$ $[ED_5] = 1.1444$ 4.1.14.Cytochrome P450-ER (PB Day 5) $[ED_5] = [ED_0] + kt$ $[ED_5] = 3.0319$ 4.1.15.NADPHCCR-ER (PB Day 5) $[ED_5] = [ED_0] + kt$ $[ED_5] = 0.0159$ 5. Level 5 Complexity **5.1. Relative Change per Average Cell Rules** 5.1.1. CCC Equation $CCC_{(t_i)} = C(t_i) \times \frac{W_L(t_i) - W[EHS(t_0)]}{W_L(t_0) - W[EHS(t_0)]}$ 5.1.2. MD equation (CytoP450) $S_{total} = MD \times U_{total}$ Control Day 0: $S_{er,liver} = 58.87$ m^2 Control Day 0: 100% PB Day 0.67: $S_{er,liver} = 97.61$ m^2 PB Day 0.67: 178% PB Day 2: $S_{er,liver} = 91.39$ m^2 PB Day 2: 167% PB Day 5: $S_{er,liver} = 108.97$ m^2 PB Day 5: 199% 6. Level 6 Complexity **6.1. Absolute Change per Organ Rules** 6.1.1. CCC Equation $CCC_{(t_i)} = C(t_i) \times \frac{W_L(t_i) - W[EHS(t_0)]}{W_L(t_0) - W[EHS(t_0)]}$ 6.1.2. MD equation (for CytoP450) $S_{total} = MD \times U_{total}$ Control Day 0: $S_{er,liver} = 58.87$ m^2 PB Day 0.67: $S_{er,liver} = 97.61$ m^2 PB Day 2: $S_{er,liver} = 91.39$ m^2 PB Day 5: $S_{er,liver} = 108.97$ m^2
Conclusions: In hepatocytes, exposure to PB significantly increases the surface area of ER and activities of its membrane-bound marker enzymes. The primer explains how to convert the results in this column to those in the complexity column.	Conclusions: In hepatocytes, PB triggers events across 6 levels of complexity. To detect and explain how these events produced a biological change, roughly 100 equations were needed.

Figure 7.34 When viewing the same original data under two different theory structures, we can see two different versions of biology. Reductionism offers a simplified picture, whereas complexity delivers details.

7.8.2 Visual Summary of the Phenobarbital Study

By copying biology's rules, this is what we can now say about the complexity of a hepatocytic response to the drug phenobarbital (PB); (Figures 7.34 to 7.37).

In Response to PB, Hepatocytes Change their ER Membrane Recipe

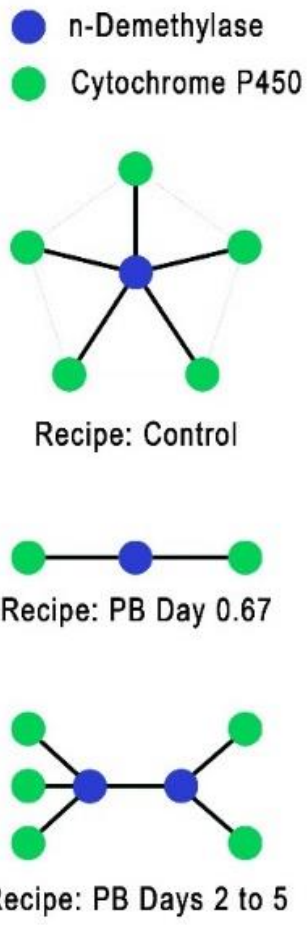


Figure 7.34 Hepatocytes change by changing the ratio recipe of its ER membrane-bound marker enzymes. For example, the ratio goes from 1 unit of n-Demethylase activity to 5 units of Cytochrome P450 activity [1 to 5 (Control Day 0)] to 1 to 2 (PB Day 0.67) and then 2 to 5 (PB Days 2 to 5); Original data adapted from Stäubli et al., 1969.

In Response to PB, Hepatocytes Change the Synthetic Rates of their ER Marker Enzymes

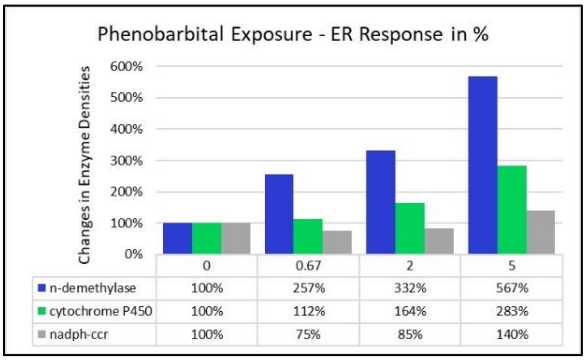


Figure 7.35 Hepatocytes change the recipe for their ER membranes by changing the rates at which new enzymes are added to the membranes. By using a synthetic rate for n-demethylase faster than the one for cytochrome P450, the enzyme ratio shifts from 1:5 to 2:5 (Figure 7.34); Original data adapted from Stäubli et al., 1969

In Response to PB, Hepatocytes Change their ER Membranes According to the Structure-Function Change Rule

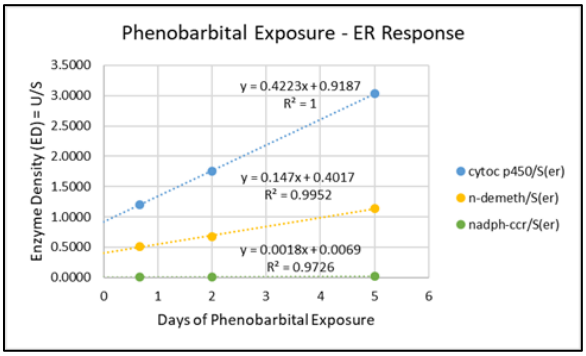


Figure 7.36 In response to phenobarbital, hepatocytes remodel their ER membranes by increasing their enzyme densities – the enzyme packing densities. This produces a membrane recipe specific to phenobarbital at a specific dose (Original data adapted from Stäubli et al., 1969).

In Response to PB, Hepatocytes Change the Rate Constants of specific membrane-bound marker enzymes of the ER. These changes follow zeroth order kinetics.

For PB days 0.67 to 5:

$$[\boldsymbol{A}] = [\boldsymbol{A}]_{\boldsymbol{0}} \pm \boldsymbol{kt}\ .$$

A rate constant defines the rate of change.

In Response to PB, Hepatocytes Change the Total Amount of Drug-Metabolizing ER Membrane in the Liver

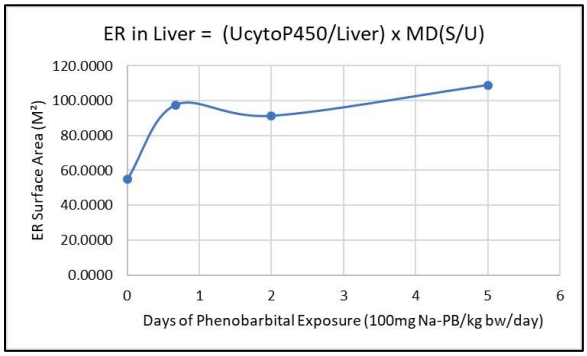


Figure 7.37 It takes 2 days of exposure to PB for the hepatocytes to come up with the new PB recipe for its ER. They then produce more of the PB membrane but continue to increase the concentration of the marker enzymes over time (detected as increasing enzyme densities); Original data adapted from Stäubli et al., 1969.

7.9 Changes in Average Hepatocytes

If we use $1.93x10^9$ as the number of hepatocytic nuclei in the liver (Marcos et al, 2006), we can calculate mashup values for average mononuclear hepatocytes (Figure 7.38).

TOTAL LIVER TO AVERAGE HEPATOCYTE				
LIVER				
PB	U/L			S/L
Days	dem	cytocP450	nadphccr	s er/L
0	11.06	58.88	0.63	54.87
0.67	50.54	117.66	0.83	97.61
2	61.24	160.66	0.88	91.39
5	124.70	330.37	1.73	108.97
AVERAGE HEPATOCYTE				
PB	U/AV HEP			S/AV HEP
Days	dem	cytocP450	nadphccr	s er/av hep
0	5.73E-09	3.05E-08	3.25E-10	2.84E-08
0.67	2.62E-08	6.10E-08	4.32E-10	5.06E-08
2	3.17E-08	8.32E-08	4.57E-10	4.74E-08
5	6.46E-08	1.71E-07	8.98E-10	5.65E-08

Figure 7.38 The figure includes data for the total liver and for average hepatocytes, wherein (dem) stands for n-demethylase, (cytocP450) for cytochrome P450, and (nadphccr) for NADPH cytochrome c reductase. (Original data adapted from Stäubli et al., 1969 and Marcos et al., 2006).

7.10 Unfinished Business

To study cells biochemically, we homogenize organs or tissues and then apply differential centrifugation to concentrate and isolate the various parts of the cells. Using analytical differential centrifugation, de Duve and his colleagues introduced balance sheets as a way of keeping track of how much of what was going where and what was getting lost in the process. This approach minimizes ambiguity by establishing ground rules for collecting, quantifying, and interpreting fractionation data.

However, most studies of the ER do not include balance sheets, preferring instead to focus exclusively on the microsomal fraction wherein most of the ER membranes are to be found. Since Stäubli et al. (1969) took this route, we need to determine what effect this might have had on the results presented herein. To do this, we'll use a standard balance sheet to see how much of the ER membranes is expected to end up in the microsomal fraction and determine the extent of the contamination coming from other cell membranes (i.e., plasma membrane (PM), inner (IMIM) and outer (OMIM) mitochondrial membranes. We can predict these membrane surface areas from marker enzyme activities in control animals using the equations given in Figure 2.4.

7.10.1 Applying Analytical Fractionation (Control Liver)

Figure 7.39 includes marker enzyme activities for the ER (G-6-Pase), plasma membrane (5'nucleotidase), outer mitochondrial membrane (monoamine oxidase), and inner mitochondrial membrane (cytochrome oxidase), as reported by Bolender et al., (1978). The equation used to calculate the membrane surface area from a marker enzyme activity is included for each organelle compartment, along

with the biochemical and morphological recoveries. Recoveries were calculated by comparing the value of the homogenate (extract (E) + nuclear (N) fraction) to the sum of the fractions (N (nuclear) + heavy mitochondrial (M) + light mitochondrial (L) + microsomal (P)) - as described by de Duve (1964, 1974). The supernatant fraction was not included in the recovery calculation because it doesn't contain any membranes.

ER		**S=0.1677xU**
Control	U/g Liver	S/g Liver
Fractions	G-6-Pase	ER (m^2)
N	4.256	0.714
M	3.639	0.610
L	2.435	0.408
P	14.740	2.472
SUM	25.070	4.204
Recovery	95%	91%

PM		**S=0.0381xU**
Control	U/g Liver	S/g Liver
Fractions	5'NUC	PM (m^2)
N	5.080	0.194
M	3.400	0.130
L	0.850	0.032
P	4.390	0.167
SUM	13.720	0.523
Recovery	95%	87%

OMIM		**S=1.5848xU**
Control	U/g Liver	S/g Liver
Fractions	MAO	OMIM (m^2)
N	0.133	0.211
M	0.353	0.559
L	0.020	0.032
P	0.087	0.138
SUM	0.593	0.940
Recovery	94%	93%

IMIM		**S=0.1467xU**
Control	U/g Liver	S/g Liver
Fractions	CY OX	IMIM (m^2)
N	4.120	0.604
M	11.100	1.628
L	0.700	0.103
P	0.460	0.067
SUM	16.380	2.403
Recovery	83%	83%

Figure 7.39 The figure includes recoveries for the marker enzymes and their associated membranes calculated using the equations given in Figure 2.4. The recoveries for the biochemical data are slightly lower than previously reported (Bolender et al., 1978) because the soluble enzyme activities in the supernatant (S fraction) were not included. The membrane-bound marker enzymes are paired with their respective organelle compartments.

7.10.2 Opening the Microsomal Black Box

We will use the data in Figure 7.39 to look inside the microsomal black box and ask a few questions.

To what extent is the ER content of the microsomal (P) fraction contaminated by other membrane organelles?

Figure 7.40, which answers the question for the control animals, shows that the microsomal fraction in this study contains 87% ER, 6% PM, 5% OMIM, and 2% IMIM. For the phenobarbital treated animals, the distribution of the membrane contamination is unknown.

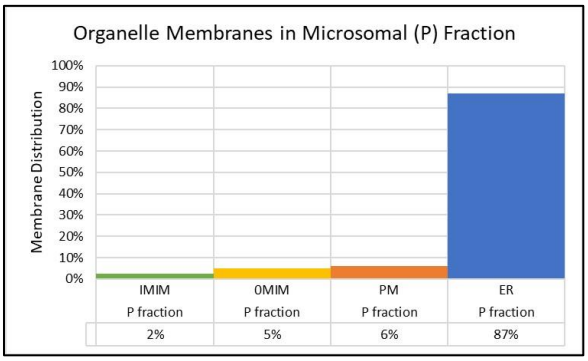


Figure 7.40 In the microsomal (P) fraction of the controls, the ER membranes have a 13% contamination coming the mitochondrial and plasma membranes. This adds roughly 13% more protein to the fraction, which would diminish the specific activity of the ER marker enzyme assay by adding roughly a 13% protein contaminant (Original data adapted from Bolender et al., 1978). Since enzyme activities are related first to a mg protein and then to a

gram of liver, similar contaminations must exist in other microsomal studies as well. Moreover, the distribution of the protein contamination across the fractions in the control may not be the same for the experimental fractions.

Next Question. What percentage of the ER membranes end up in the microsomal (P) fraction?

For the control animals, 59% of the ER sediments in the microsomal fraction, with lesser amounts being distributed in the L (10%), M (15%), and N (17%) fractions (Figure 7.41). For the phenobarbital treated animals of the Stäubli study, the distribution of the ER membranes and marker enzyme activities in the fractions is unknown.

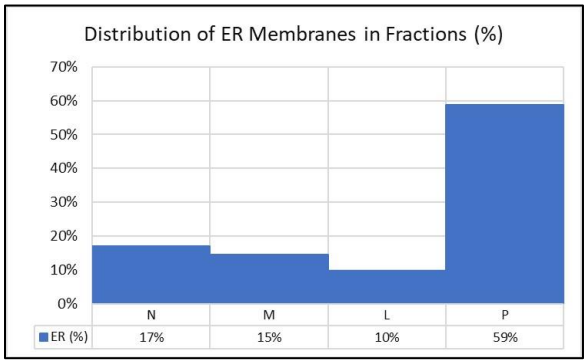


Figure 7.41 The ER occurs in all the membrane containing fractions, with 17% going to nuclear (N), 15% to the heavy mitochondrial (M), 10% to the light mitochondrial (L), and 59% to the microsomal (P) (Original data adapted from Bolender et al., 1978).

What happens to the structure-function change equations of Figure 4.24 if we move all the ER marker enzyme activity (n-demethylase) into the microsomal (P) fraction (using the distribution given in the control profile of Figure 7.41), recalculate, and plot the revised enzyme densities?

We get the results shown in Figure 7.42. The enzyme densities continue to produce structure-function change equations, but they have different slopes and y intercepts.

	ED	ED	ED
Days	nadph-ccr/S(er)	cytoc p450/S(er)	n-demeth/S(er)
0	0.0194	1.8189	0.3418
0.67	0.0145	2.0430	0.8776
2	0.0163	2.9796	1.1359
5	0.0270	5.1388	1.9396

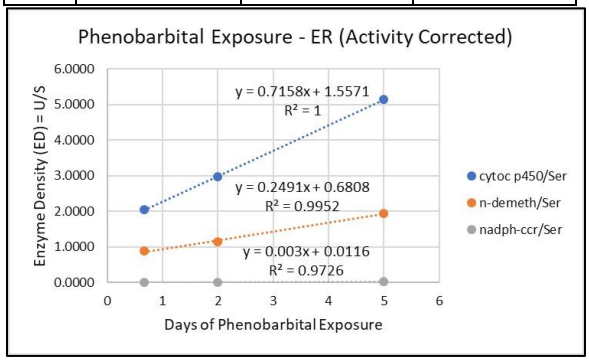


Figure 7.42 When the control and experimental values are corrected to the total liver values, the same pattern as seen previously for the uncorrected values (Figure 4.25) remains unchanged (Original data adapted from Stäubli et al., 1969 and Bolender, et al., 1978).

Do the equations in Figures 4.22 and 7.42 detect the same relative amount of change?

Yes. Figure 7.43 shows that the enzyme densities detect the same amount of change when related to either the ER in the microsomal fraction or to the liver.

	Microsomes (P)	Liver	Same for Both
	ED	ED	% Change
Days	**n-demethylase**	**n-demethylase**	**n-demethylase**
0	0.2017	0.3418	100%
0.67	0.5178	0.8776	257%
2	0.6702	1.1359	332%
5	1.1444	1.9396	567%

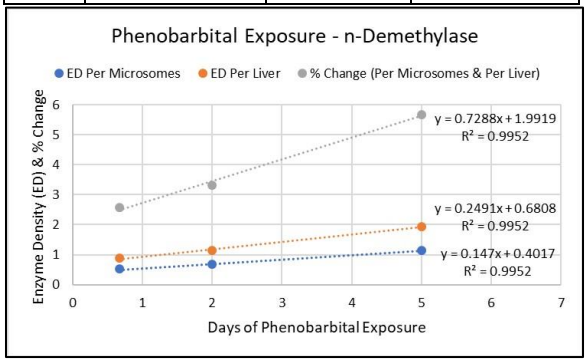


Figure 7.43 When the enzyme densities are expressed as a percentage of the controls (1 = 100% for the microsomes or liver), they identify the rate at which the ER membranes become enriched with the marker enzymes. Both sets of

enzyme densities detected the same percentage change (Original data adapted from Stäubli et al., 1969 and Bolender et al., 1978).

Notice in Figure 7.43 that the enzyme density and percentage change curves plot the same data but produce clearly different slopes and y intercepts. When looking for a one to one relationship, this distinction between relative and absolute change becomes important.

In short, what is the microsomal black box telling us?

Since multiple variables are in play (Figures 7.39 to 7.43), it seems fair to suggest that our chances of reproducing published microsomal studies are remote.

But how do we deal with such a shortcoming?

Let's first try to understand the microsomal problem a little bit better and then figure out how we can solve it.

7.10.3 Untangling the Variables in the Microsomal Black Box

Microsomes represent a methodological construct subject to the inconsistencies produced by differences in assay protocols, fractionation techniques, data references (mg protein), and animal variation. Let's see what we can learn about this construct by asking a few more questions.

How do the control values for a given enzyme assay (n-demethylase) compare across publications?

Figure 7.44 indicates that they can differ considerably - even by orders of magnitude. Recall that we found a similar pattern earlier using data collected from the brain (Figure 4.15).

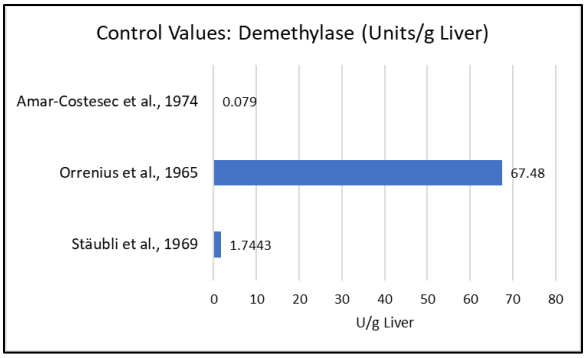


Figure 7.44 Three independent estimates for the activity of n-demethylase in the microsomal fraction of control livers produced widely different values.

Is the protein content of the microsomal fraction – the data reference for enzyme assays - consistent across publications?

No, not according to the studies shown in Figure 7.45. Once again, we can find an order of magnitude spread of the data.

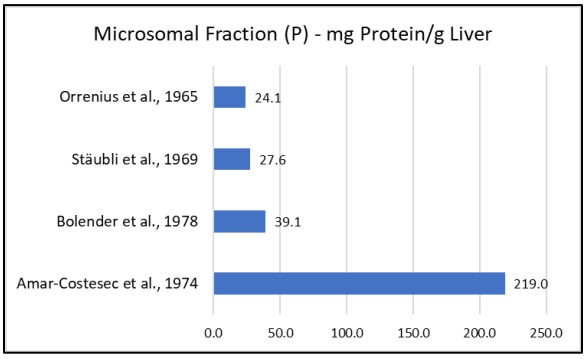


Figure 7.45 The plot illustrates control values for the mg protein reference assayed in the microsomal (P) fraction.

Considering the methodological inconsistencies identified in the literature by Figures 7.40, 7.41, 7.44, and 7.45, the design and application of the microsomal model being used to detect biological changes seems poorly prepared to pass a reproducibility test. Each of these four figures makes the same point. If we want reproducibility, then it should be part of our experimental design.

Is the phenobarbital study reproducible?

Since the phenobarbital study of Stäubli et al. (1969) duplicated an earlier one by Orrenius et al. (1965), we can do a mashup and calculate a second set of enzyme densities based on an independent set of microsomal assays for n-demethylase.

When we plot the results of Orrenius and Staubli papers (Figure 7.46), we find that the amount of change in the enzyme densities differs by a factor of 2. The answer to the question is yes and no. The rule is reproducible but not the changes it detects. Why? The assay and fractionation methods are not the same.

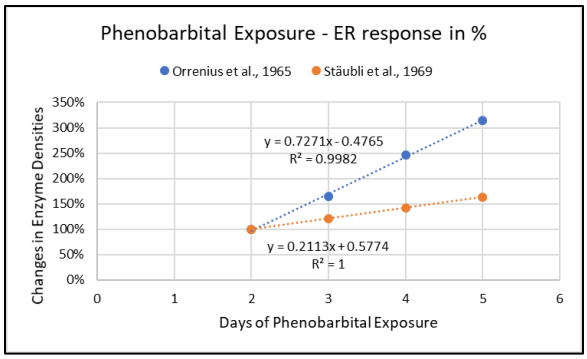


Figure 7.46 The same experiment – based on microsomes – detects different enzyme densities (n-demethylase/ER), but the data of both studies are subject to the same underlying structure-function change rule. Note that the enzymes densities of the Orrenius study represent mashups, using ER surface areas borrowed from the Stäubli study change (Original data adapted from Orrenius et al., 1965 and Stäubli et al., 1969).

Let's review. We can expect that microsomal studies will be difficult to reproduce because several variables are changing at the same time. Key variables include:

1) Changes in the percentage of the total ER membranes in the microsomal fraction (Figure 7.41).
2) Changes in the enzyme densities (Figure 7.42).
3) Changes in the contamination of the microsomal ER coming from other membrane organelles (and from other cell types), which adds a protein contamination (Figure 7.40). Recall that enzyme assays are related to a mg of protein, which is measured in each fraction.
4) Differences in the fractionation methods used to collect the microsomes influence the percentage of ER in the microsomal fraction (Figure 7.41).
5) Differences in the methods being used to assay n-demethylase activity detect different amounts of change (Figure 7.44).
6) Differences in the methods and standards used to assay protein influence whatever we relate to the mg protein reference. This includes both enzyme activities and the enzyme densities derived therefrom (Figure 7.45).
7) Differences in enzyme assays produce different enzyme densities (Figure 7.46).

By opening the black box and looking inside, we can see what the experimental model used to detect biological changes with microsomal data collapses under the weight of at least seven confounding variables. Effectively, this puts the reproducibility requirement well beyond reach. How do we solve this problem? Design a new experimental model, one that embraces reproducibility and avoids troublesome variables.

7.10.4 Opening the Phenobarbital Black Box

Figures 7.23 and 7.26 suggest that the relationship of two of its ER marker enzymes (n-demethylase to cytochrome P450) becomes optimal at day 2 of phenobarbital treatment and persists thereafter.

Since we now suspect that the amoeba uses some form of linear programming to optimize its travelling salesman problem (Zhu et al., 2018), how might the hepatocyte optimize a

solution to its problem of producing drug-metabolizing membranes?

Recall that a linear programming problem involves the relationship between variables, which are proportional to one another. Such a problem can be solved graphically as the intersection of two linear equations subject to constraints.

Although the two enzyme density curves displayed in Figure 7.47 intersect (x = -1.8779, y = 0.1256), the negative x value does not fit the typical linear programming model. However, if the intersection represents a connecting value, and if the ratio of the marker enzymes is optimized ("via biology's math"), then it might explain why all the data points fit the curves as well as they do.

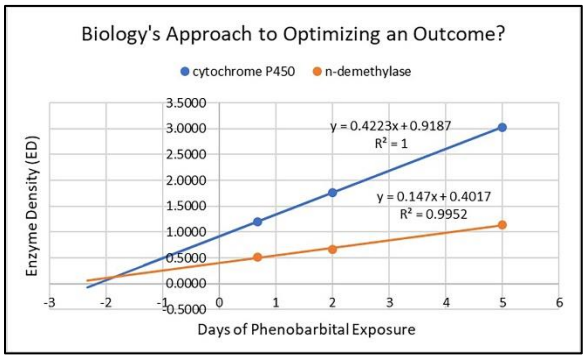


Figure 7.47 The two enzyme density curves intersect at x = -1.8779 and y = 0.12564.

The hepatocyte uses two strategies when metabolizing a drug. It increases the concentration of the drug-metabolizing enzymes per unit of ER membrane and/or increases the total amount of ER membrane in the cell. What we don't know is the relationship of the drug dose to the enzyme density and to the total cellular ER.

What happens, for example, to the curves in Figure 7.47 and to the total liver ER when we double or half the daily dose of phenobarbital? How would such data fit into the pattern shown in Figure 7.47? In effect, these are mathematical questions that we could ask biology to answer.

7.11 Optimize the Structure-Function Change Experiment

Now let's redesign the phenobarbital experiments of Stäubli et al., (1969) and Orrenius et al., (1965) but this time leave out most of the troublemakers.

7.11.1 Optimizing the Experimental Design

We begin by asking "What do we want?" We want an optimal experimental design, one that minimizes input (time and resources) and maximizes output (understanding, usefulness, and reproducibility). The following protocol should get the job done:

1) Use stereology to collect stereological data from intact livers. Apply the Weibel-Paumgartner corrections.
2) Use an abbreviated form of fractionation to collect biochemical data. This involves collecting just two fractions (extract (E) and nuclear (N)), assaying marker enzymes in both of the fractions, and expressing the results per homogenate (H). Recall that H = E + N.
3) Integrate and analyze the structure-function data as described in the primer.
4) Publish the details of the enzyme assays in minute detail.

Figure 7.48 illustrates a minimum data set designed to study complex changes in hepatocytes. For the phenobarbital study, drug-inducible enzymes (e.g., n-demethylase, cytochrome P450) would be added to this basic protocol.

OPTIMAL EXPERIMENTAL MODEL				
Function			Structure	F/S
Marker Enzyme			Membrane	
U/g			S/g	U/S
E	N	H = E + N		ED = U/S
Glucose-6-Phosphatase			ER	
22.620	4.256	26.876	4.600	5.843
Cytochrome Oxidase			IMIM	
15.53	4.12	19.65	2.883	6.815
Monoamine Oxidase			OMIM	
0.506	0.133	0.639	1.013	0.631
5' Nucleotidase			PM	
10.75	5.08	15.83	0.603	26.267

Figure 7.48 An optimal experimental model includes a collection of membrane-bound marker enzymes connected to their morphological locations and expressed as enzyme densities – estimated in the most efficient and reliable way. The figure illustrates a minimal data set for each control and experimental time point (Data adapted from Bolender et al., 1978).

This optimal experimental design, which juxtaposes total liver structure to total liver function, mitigates the effects of the troubling variables by eliminating the need to collect microsomes.

7.11.2 Why is Optimization Needed?

Figure 7.49 lists the key features of the optimal experimental model.

FEATURES OF AN OPTIMAL EXPERIMENTAL MODEL			
	Items	Present	Absent
1	Concentration Trap		X
2	Membrane Contamination		X
3	mg Protein Contamination		X
4	Stereological Volume Distortions	X*	
5	Section Thickness Biases		X
6	Differences in Enzyme Assays	X*	
7	Differences in Fractionation Methods		X
8	Recovery Problems		
9	Change	X	
10	Reproducibility	X	
11	Precision	X	
12	Accuracy	X	
13	Prediction	X	
14	Diagnosis	X	

Figure 7.49 An experimental model is distinguished by what is does and does not do. Moreover, it should include all the items in the package (9 to 14).

Note that two items (*) in Figure 7.48 continue to thwart the goal of optimization – stereological volume distortions and inconsistent enzyme assays. Suggestions for dealing with the stereology problem exist elsewhere (Bolender, 2013), but we still must deal with the assay problem.

If we assume that enzyme assays will continue to show important differences across laboratories, then standardization based on control values may become a workable option. We can do the following. Translate the enzymatic data into relative changes, identify a standard (a control value averaged across the literature), and multiply the relative changes by the control standard. This option can be applied to both protein and enzyme assays.

Will standardization work for the microsomal assays of the two phenobarbital studies? Figure 7.50 says no. Why? When dealing with the microsomal fraction, at least four variables are in play: protein content, enzyme assay, membrane contaminations, and the percentage of the ER in the microsomal fraction. Taming one variable helps, but reproducibility requires taming them all. This is what the optimal model is designed to. It eliminates the unavoidable shortcomings of microsomes by substituting homogenates.

n-demethylase/S(er)		
	Orrenius-1965	Staubli-1969
Days	ED	ED
0	1.1149	1.1149
1	1.4085	2.6650
2	1.8410	3.0923
3	2.7837	3.4681
4	3.9289	3.8011
5	4.7059	4.0984

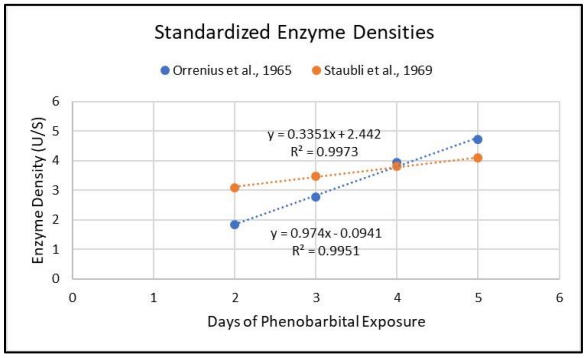


Figure 7.50 Compared to Figure 7.46, standardizing the enzyme densities (n-demethylase/S(er)) improved the results but failed to fix our reproducibility problem. Only one of the four time points (PB day 4) was reproducible. Too many variables are still in play.

7.12 What Did you Learn?

What can you do now that you couldn't do before reading the primer?

You can read a research paper, look at the data types, and evaluate both the strengths and weaknesses of the authors' results and conclusions.

You now know how to look at the same data operating under two different theory structures. When looking at the human amygdala through the eyes of reductionism, for example, you can see data being heavily influenced by biological variation (Figure 4.15), whereas complexity theory lets you view the same data when they are obeying a biological rule (Figure 4.16). Moreover, you discovered that stereological data tell different stories before and after being corrected for methodological biases (Figures 3.16, 3.17, 3.19, 5.3).

You know how to use quantitative approaches to hunt for complex patterns in the literature (Chapter 4) and understand that local (per individual) and global (per population) patterns are subject to the same rules and the same test of reproducibility (Figures 4.15, 4.16).

You know that problem solving involves finding and following clues. The current reproducibility crisis, for example, provides a helpful clue by calling attention to a fundamental flaw in our approach to collecting data. Unwittingly, our experimental methods routinely transform pure biological data into artificial data types by mixing two sources of variation (methodological and biological) - each of which exhibits its own inter- and intra-variation. This explains why our attempts to deliver reproducibility are - for all practical purposes – designed to fail. Since you know that it's also possible to design approaches that succeed, the reproducibility problem becomes manageable. You can minimize the influence of the artificial data types by optimizing the experimental design (Figures 7.48 and 7.49) and by using a first principles approach to demonstrate that the same biological rule is being applied repeatedly (Figures 4.2, 4.4, 4.16, 4.18, 4.22, 4.23, 7.26, 7.46 and 7.50).

You now understand why complexity in biology depends importantly of the relationship of structure to function and how rule-based approaches to the basic and clinical sciences avoid much of the mischief created by simplifying biology, which reduces it to a catalogue of disconnected parts (Chapter 7).

You know that a biological change involves the orchestration of many interacting rules, all of which become obscured when experiments are confined to black boxes. You know how to open such boxes, look inside, and view the

underlying complexity of a biological change. You can do this because you have access to biology's package of skills and know when, where, and how to discover and apply rules by reasoning with equations.

You have seen how biology repeatedly incorporates ratios into its rules and that most rules are defined by the relationship of structure to function.

You have discovered that steps can be taken to upgrade a descriptive science to a hard science by deriving it from first principles. Many first principles of physics and chemistry - discovered using a theory structure based on reductionism - are already etched in stone. Biology is less forgiving. We tried reductionism only to discover that biology obeys rules etched both in stone and in genes. This requires a double dose of theory structure – reductionism to take biology apart and complexity theory to put it back together. By identifying first principles related to genes and their phenotypes, biology's rules become a primary route to discovery in the life sciences. Solving biology is now akin to solving physics and chemistry. It's done mathematically with equations.

Fortunately for us, biology has already solved itself mathematically and all you have to do is follow the clues and copy biology.

CHAPTER 8

HOW TO SOLVE BIOLOGY

The primer describes how we can approach solutions to a wide range of challenging problems by moving our published data from journals to databases and by upgrading our reductionist model to one of complexity. As a part of this process, we have learned how to recruit biology as our main problem solver by deferring to its rules and first principles. This allows us to peer into the black boxes created by reductionism and to gain access to biology's highly prized package of skills.

But how do we use these newfound abilities to solve biology? By upgrading the biology literature to a complexity model, we now have access to large amounts of highly connectable information. This resource can be used to address new generations of problems in biology by assembling mashups. Recall that a mashup is a derivative work created by combining individual pieces to create something new and more useful. Its main characteristics often include combination, visualization, and aggregation.

The purpose of this chapter is to illustrate this process by going through the major steps of assembling a typical mashup – using the liver as an example. The process of solving biology is the same as it is for physics and chemistry – use the theory structure as a guide to identify and verify equations that allow us to discover, innovate, and solve problems. Such is the nature of a science based on rules derived from nature.

8.1 LIVER MASHUP

8.1.1 SETTING THE GOALS

In short, we would like to explore the rule-based relationships of liver phenotypes to the upstream rules (genetic recipes) responsible for their design - across a wide range of experimental and clinical settings. Specifically, where, when, and how are the genetic recipes assembled and what steps and requirements are involved in passing information back and forth between genotypes and phenotypes? The answers will most likely come from published data in the form of ratios, algorithms, rules, equations, and networks of mathematically defined interactions

Currently, we view genotypes as collections of genes, whereas biology may have a more expansive view. Why? Genes make up only about 2% to 3% of the DNA in our genomes. Might we be missing something – something just as important? Will progress in understanding the information coded in our DNA be advanced by assembling mashups that combine the data, algorithms, rules, and equations of cell and molecular biology into search and discovery platforms? What are we likely to discover? The likely insight to come from such questions is the conclusion that solving the genotype begins by solving the phenotype.

The incentive for such an exercise would be to produce an information resource appropriate to the task of integrating artificial intelligence, change, diagnosis, prediction, and reproducibility into the core disciplines of the

basic and clinical sciences. Here, the argument becomes one of direction and timing. By reverse engineering biology, forward engineering inherently becomes a safer and more productive enterprise. Now that CRIPSR has opened Pandora's box, a comprehensive approach to genetic modification will at some point become a priority issue - if it hasn't already.

8.1.2 Overview

Our goal here is to reconstitute the structural and functional complexity of the liver – by rule – from a large collection of parts available to us in the literature. To this end, we can begin the process by interconnecting structure-function equations such that the output variable of one equation becomes the input variable of another. This will allow us to move information freely within and across multiple levels of complexity. Using the phenotype equations included in chapter 7, we already have the wherewithal to use seed values and lookup tables to generate quantitative phenotypes for the molecules, organelles, and cells that define the liver.

By deferring to biology's rules and equations generated from parallel complexities, published data from a wide range of otherwise incompatible publications can contribute to putting the liver back together. In effect, by marshalling contributions from thousands of investigators, the resulting phenotypes become representative of biology and compatible with a wide range of research programs.

8.1.3 Primary Sources of Liver Data

Since mammalian livers within and across species are likely to share many similar genes and rules (Alberts et al., 2014, Lodish et al., 2016), the mashup can assimilate data coming from multiple species. We already know, for example, that humans and rats can use similar rules to organize cytoplasmic membranes in hepatocytes (Bolender, 2017).

8.2 Mashup Procedure (Liver)

The long-term goal of the mashup procedure is to assemble quantitative phenotypes for the liver that can not only identify, interpret, and explain the rules governing gene expression, but also to allow diagnosis, prediction, and *in situ* simulations. Such an approach is consistent with the concept of a universal biology database that allows us to treat data collected from the literature as one big experiment.

We will start the mashup process with hepatocytic ER, continue the work up with other membrane organelles, add the particle organelles (peroxisomes and dense bodies), and repeat the entire process for Kupffer, fat-storing, and endothelial cells.

8.2.1 Step 1 – Locate Data by Reading Papers

By any measure, accessing data remains the greatest challenge because many research publications continue to exist behind paywalls. Recall that when building even a modest database, which involves collecting data from about 5,000 studies, an investigator can expect to browse roughly 15,000 papers.

8.2.2 Step 2 – Standardize Data

Given conventions well-established in the biology community, absolute values and concentrations represent the most widely published data types. The goal here is to

generate data tables from each paper to speed data entry and to minimize errors. Data are standardized one paper at a time.

8.2.3 Step 3 – Generate Primary and Derivative Literature Databases

Use relational databases to manage published data. This includes generating derivative databases capable of targeting specific goals.

8.2.4 Step 4 – Generate Universal Biology Databases

By translating the data stored in the literature database into mathematical markers and connection ratios, for example, they become primed to find quantitative patterns in the literature. Such patterns, which derive from quantitative relationships, can serve as effective diagnostic and predictive tools. Moreover, they provide clues as to when, where, and how key events are happening.

8.2.5 Step 5 – Generate Rule-Based Equations

Modular data sets connected by equations combine to detect and replicate complex changes occurring in hepatocytes. Identify the equations used to describe events occurring in complexity levels 2 to 6, and store them in tables and databases suitable for distribution.

By taking a modular approach to reconstructing biological events, equations and data can supply local and global information, be combined to form simulators, integrate published data, predict missing values, and generate diagnostic and predictive patterns.

One of the unexpected surprises to come from the experience of building a biology database is that published data are sparse and often difficult to interpret as a complexity. Fitting existing data to equations and then using them to predict missing information offers a workable solution to this limitation.

8.2.6 Step 6 – Run Worksheet Simulations

Fundamental to creating large scale mashups is an ability to generate and synchronize multiple linear pathways – structurally and/or functionally. Modules, identified as the equivalent of phenotypic codes, can provide workable solutions to both standardization and connectivity across publications. In effect, modules become an essential building block of the mashup. For example, enzyme densities serve as modules that combine to generate structure-function equations capable of solving a wide range of problems.

8.2.6.1 Predict Biochemistry from Biochemistry

Phenotypic codes for biochemical data identify - as ratios - the relationship of one enzyme activity to another as a linear string – according to the ratio chain rules. This defines a chemical stoichiometry as the relative amounts of enzyme activities in a connected set of ratios making up a string. Such an arrangement characterizes the biochemical phenotype as it exists under control and experimental conditions. Using this rule, we can generate data from seed values and aggregate large data sets across a wide range of experimental settings (Figures 7.15 and 7.18). Here the objective is to translate point data into well-springs of new information.

8.2.6.2 Two Way Predictions

Since a similar stoichiometry exists among the structural components of cells and tissues (Figure 7.16), morphological strings can be generated and used in a similar manner. Furthermore, we can use ratios to connect the organelle compartments of four cell types (hepatocytes, Kupffer cells, fat-storing cells, and endothelial cells), which together make up the liver parenchyma (Figure 7.17).

8.2.4.3 Predict One from the Other: Structure → Function and Function → Structure

Structure-function equations – based on enzyme and membrane densities - have the capacity to predict biochemistry from morphology and morphology from biochemistry (Figure 7.15). These equations, which can be generated for control (Figure 7.15) and experimental settings (Figure 2.5) supply missing information, fingerprint phenotypic states, verify reproducibility, and produce a new generation of complex data types.

Moreover, we now have the option of using enzyme and membrane densities to generate matching data for papers reporting just morphological or biochemical results.

8.2.7 Step 7 – Calculate Enzyme and Membrane Densities for Cell Organelles

Using the worksheet models shown in Figures 2.1 (structure-function rule), 7.15 (control), and 2.5 (change), tables of structure-function equations can be accumulated and used to define liver phenotypes in a variety of settings.

8.2.8 Step 8 – Calculate Rate Constants for Biochemical Changes in Organelles

When simulating portions of the metabolic chart, to which the liver contributes generously, the most problematic variables become the rate constants, which come from *in vitro* estimates. Enzyme densities, which relate changes in enzyme activities to a standard unit of membrane surface area (1 m^2) - instead of the usual mg of protein or gram of liver – offer the advantage of estimating the rates of biochemical changes *in situ*.

8.2.9 Step 9 – Calculate Absolute Values for Cell Compartments

The equations used in Figure 7.31, which calculate membrane surface areas from membrane densities and units of enzyme activity, offer a workable solution to the absolute data problem created by the hierarchy equations of stereology. A similar approach works for biochemical data (the equations in Figures 7.15 to 7.18).

8.3 Online Resources

8.3.1 Websites

The Enterprise Biology Software Project introduces websites in step with the evolution of the project:

2001: enterprisebiology.com

- Yearly Progress Reports (1 to 16)
- Software packages

2016: playingcomplexitygames.com

- eBook: ***Playing the complexity Game with Biology***

- Considers strategies for interacting with biology as a complexity
- Includes picture-based summaries
- Offers databases, directions, worksheets, and templates
- Catalogues yearly reports (2001 to 2018)

2019: solvingbiology.com

- eBook: ***Solving Biology: A Primer***
- Serves as a beginner's guide to solving biology as a complexity
- Includes picture-based summaries

8.3.2 Yearly Reports

The EBSP reports and support material were originally distributed on CD or DVD to contributing authors located in more than 45 countries. The support material included literature databases, associated computer programs, and a narrative consistent with the view that discovery becomes the natural consequence of studying biology as a complexity. Similar resources are available online.

REPORTS

2001: Aggregating Research Data

2001: Background Courses: biology, stereology

2002: Patterns in Relational Databases

2003: Simple and Complex

2004: Data to Equations

2005: Puzzles/Universal Biology Databases 1.0

2006: Universal Biology Databases 2.0

2007: Universal Biology Databases 3.0

2008: Universal Biology Databases 4.0

2009: Systems Biology Two

2010: Organism Codes

2011: Mathematical Mapping

2012: Mathematical Markers

2013: Complexity Games

2014: Big Data

2015: Disorder-Order

2016: Precision-Accuracy-Reproducibility

2017: Biological Homogeneity

2018: Biological Change

CHAPTER 9

THEORY OF BIOLOGICAL COMPLEXITY[2]

9.1 INTRODUCTION

Complexity in biology involves two seemingly inconsistent layers of information: rules and adaptability. The rules layer exhibits little or no variation, whereas the adaptability layer allows widespread variation. To us, one appears statistically quiet, the other noisy.

The biology literature, which houses a vast collection of parts data, focuses almost entirely on data coming from the adaptability layer. Consequently, the current iteration of the literature operates largely in the absence of biology's rules. A central challenge of the primer was to show how we can move the parts data from the adaptability layer back onto the rules layer where they normally occur in nature. This was accomplished by recovering the missing connections and complexities.

Recall that rules – like genes – tend to be stable and widely shared within and across species, whereas the parts of an individual vary in response to local conditions as the need arises. This helps to explain why we rely on statistics to detect simple changes in parts (high variability), but patterns, rules, and equations to capture complex changes (low variability).

The theory of biological complexity encourages us to ask the most telling question. Why does biology operate by rule, but not the science of biology? This fact seems sadly amiss and counterproductive. The one thing we seem to be missing is a theory structure that encourages us to derive biology from first principles and the rules derived therefrom. This is how physics and chemistry operate. Why not biology?

Recall that the overarching principle of the theory structure being proposed herein states that it takes a complexity to solve a complexity. This means that to test the theory empirically we need to construct parallel complexities as close to the original as possible, relying exclusively on the rules that exist first in biology and then mirrored in our reconstituted complexities. In effect, complexity theory becomes the evolving product of copying biology.

Unfortunately, biological complexity remains largely an unfamiliar place. New rules apply, our perceptions change, and we ask and answer questions differently. Our first order of business, however, is to learn the rules of the game and then use them to assemble a more productive and insightful theory structure. This becomes an ongoing process wherein the theory unfolds in step with the discovery process. Solving biology begins by solving the phenotype, which the Primer explains how to do.

The fundamental building blocks of a biological complexity include parts and connections. Volumes, surfaces, lengths, and numbers (counts) define the basic elements of the parts quantitatively, whereas the ratios of the parts define the connections. From this simple beginning, the complexity of an organism grows as an assemblage of parts and connections cascading across the hierarchical levels of an

[2] Chapter updated from the 2014 Report (appendix III) and chapter 8: *Playing the Complexity Game, 2016.*

organism. The rules become the glue – the mathematics - that holds everything together.

Since complex data sets consist more or less of the same basic building blocks wherein all such blocks interconnect, our parallel complexities begin to mirror the original biology – at least at a beginner's level. Testing this new theory structure consists of looking for persistent patterns - locally and globally – and then using such patterns to define the rules of the game we are trying to play in collaboration with biology.

The primer encourages the reader to ask basic questions about how biology is to evolve as a science. By introducing new theory structures, we get to create new platforms for innovation and discovery, free from the many limitations imposed by entrenched, but outdated theories. The goal becomes one of coexisting with biological complexity, not ignoring it, or pretending that it doesn't exist.

Ideally, we want to use complexity theory to reconstruct biology mathematically from first principles using published data as our source of empirical proof. Thus far, the working theory structure has generated a wide range of equations capable of capturing several of biology's principles and rules. In effect, the process of redefining biology in terms of the rules and principles defining its mathematical core is now well underway. We now know that biology uses linear models based on ratios of its parts to assemble phenotypes and to change them.

9.2 A First Principles Approach

Biology – as a science – can be said to derive from first principles when it relies on the basic and established laws of nature. In the absence of such principles, biology defaults to the models and assumptions often driven by necessity or convenience.

Can we derive biology from first principles? The answer is yes. If biology can do it, we can do it too. When we bring the power of published data coming from thousands of highly skilled investigators into conjunction, the basic principles and rules of biology begin to appear almost effortlessly.

The argument for a first principle approach to biology becomes even more compelling because of the opportunities it creates. Why?

An insightful answer comes from Elon Musk. *"We normally think by analogy - by comparing experiences and ideas to what we already know - but there's a better way to innovate. I think it's important to reason from first principles rather than by analogy. The normal way we conduct our lives is we reason by analogy.* [With analogy] *we are doing this because it's like something else that was done, or it is like what other people are doing.* [With first principles] *you boil things down to the most fundamental truths…and then reason up from there."*

Musk continues: *"The benefit of first principles thinking? It allows you to innovate in clear leaps, rather than building small improvements onto something that already exists."* However, he warns us about using first principles for innovating: *"It takes a lot more mental energy."*

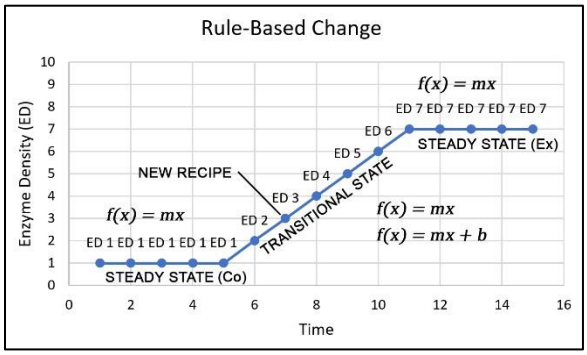
As the primer suggests, first principles become one of the many rewards to come from playing the complexity game with biology. They allow us to approach biology as a mathematical science, create universal databases from the biology literature, understand the nature of change in biology, identify widespread connectivity, work out data driven methods for clinical diagnosis and prediction, harmonize living and postmortem data, unfold the disease process, manage our current problems with

reproducibility, and profit from the untold riches of biological complexity.

9.3 First Principles

9.3.1 Principle 1: Structure-Function Rule

Two general equations define a biological change, which begins at complexity levels 2 and 3 (Figure 9.1). As the figure shows, two steady states surround an intervening transitional state. The curve identifies a string of enzyme densities (blue dots) that define the change as it occurs over time. The steady states require a single equation, whereas the transitional state requires two.



CHANGE	CONCENTRATION	RATIO OF PARTS	MEMBRANE
STEPWISE	ED	RECIPE	M^2
ED1 → ED1	↔	↔	↔
ED1 → ED2	↑	Δ	↑
ED2 → ED7	↑	↔	↑
ED3 → ED7	↑	↔	↑
ED7 → ED7	↔	↔	↔

Figure 9.1 A rule-based change involves two steady states (before and after) separated by a transitional state wherein the changes occur. The blue dots identify enzyme densities (ED) before (ED1), after (ED7) and during (ED2 to E 6) a change. Table legend: no change (↔), increase (↑), and change in recipe (Δ). Once a new recipe is established (ED2, the ratio of the parts remains constant, but their packing density increases until the new steady state is reached (ED7).

The first biological principle in play includes the structure-function rule (based on the postulates of biochemical homogeneity), which relates units of marker enzyme activity to a square meter of membrane surface area:

Structure-Function Rule

$\boldsymbol{f(x) = mx}$ where,

$U = m \times S$, and

$m = U/S = ED.$

Note that the enzyme density (ED) remains constant during the two steady states before (ED 1) and after (ED 7), but changes during the transitional period (ED 2 to ED 7).

9.3.2 Principle 2: Structure-Function Change Rule

A biological change in a membrane organelle such as the ER includes multiple events:

- Changes in the packing densities (EDs) of the biochemical constituents in the membrane,
- Changes in the ratios of the biochemical constituents (the membrane recipe), and
- Changes in the total amount of the ER membrane surface area in average cells and organs.

The changes of the transitional state adhere to the following rules:

Structure-Function Rule

$\boldsymbol{f(x) = mx}$ where,

$U = m \times S$, and

$m = U/S = ED.$

Structure-Function Change Rule

$\boldsymbol{f(x) = mx + b}$ where,

$ED = mx + b,$

$x = time, and\ b = y\ intercept$

Recipe Rule

$\boldsymbol{a(ED_1): b(ED_2): c(ED_3) \ldots x(ED_n)}$

$aED_1 bED_2 cED_3 \ldots xED_n$

Absolute Change per Organ Rule

$\boldsymbol{f(x) = ED \times S_{total}}$ where,

$U_{total} = ED \times S_{total}$

$\boldsymbol{f(x) = MD \times U_{total}}$ where,

$S_{total} = MD \times U_{total}$.

9.4 Theory of Biological Complexity

In its simplest form, the theory states that it takes a complexity to solve a complexity. We can define a biological complexity mathematically as a distinct set of elements (parts and connections) that combine to form the same rule-based patterns encountered both locally and globally. Typically, biology displays its complexity as stoichiometries based on the ratios of it parts and applies this rule to create both order and disorder.

A theory structure includes a set of guidelines for exploring biology as a complexity. The following lists summarize the goals, requirements, principles, and equations, which, when taken together chart the progress to date in assembling a new and more biologically friendly theory structure.

9.4.1 Goals

- Organize and generalize the data of the biology literature.
- Define and test a data-driven approach to the basic and clinical sciences.
- Identify mathematical patterns in biology.
- Explore biology as a rule-based system.
- Capture biology's rules empirically with equations that carry R^2s equal or close to one.
- Use published data to create parallel complexities based on rules intrinsic to biology.
- Minimize bias in experimental systems.
- Avoid the negative effects of biological variation.
- Use biological variation to detect rules.
- Optimize the definition of reproducibility.
- Remove postmortem data distortions by applying corrections based on living standards.
- Demonstrate with practical examples the effectiveness of an approach to problem solving based on empirical data guided by verifiable rules.
- Quantify biological phenotypes and use them to diagnose and predict outcomes throughout the biological hierarchy of size.
- Reevaluate traditional methods of collecting, reporting, and interpreting data in the basic and clinical sciences.
- Optimize experimental designs.
- Assemble diagnostic databases from the biology literature capable of diagnosing disorders and diseases.
- Develop methods for extracting useful patterns from large data sets.
- Identify the strategies being used by biology to create disorders.
- Identify and catalogue quantitative relationships of structure to function.
- Detect and unfold the complexity of biological changes.
- Develop a rule-based strategy for connecting phenotypes to genotypes.
- Reverse-engineer phenotypes and genotypes by rule.

9.4.2 Requirements

- Collect biological data with unbiased sampling methods.
- Express data as volumes, surfaces, length, numbers, or derivatives thereof.
- Assemble data as connected sets, consisting of ratios, mathematical markers, connection ratios, strings, and networks.
- Integrate data quantitatively within and across hierarchical levels.
- Use a universal data format – based on ratios - to organize and generalize published data.
- Operate within the bounds of complexities parallel to the ones defined by biology.
- Correct the volume distortions associated with postmortem data.
- Tune databases – by applying filters - to enable and strengthen their diagnostic and predictive properties.
- Copy biology by copying patterns, rules, equations, and algorithms.

9.4.3 First Principles (as Statements)

To derive biology from first principles, we begin by identifying rules as prospective candidates. We can assume that such rules translate into biological patterns, equations, and algorithms that define biology as a complex adaptive system. Rules that generalize become the principal candidates for first principles.

Rule 1: Biology is a complexity consisting of parts and connections.

Rule 2: Biology defines and controls its complexity by using ratios of one part to another.

Rule 3: Biology assembles complexity with strings, modules, and networks of parts connected by ratios.

Rule 4: Biology allows the same two parts to form different ratios (valences).

Rule 5: Biology allows considerable variation in the size of its parts, but minor variation in the quantitative relationship of one part to another. It maintains rule-based order, which can change during growth, aging, disease, and adaptive responses.

Rule 6: Biology uses a modular approach to construct itself and to change.

Rule 7: Biology can define complexity with modular structures, starting with two parts (a, b) with two values (x, y) and one connection (ax:by).

Rule 8: Biology consists of nested complexities, which fold and unfold by rule.

Rule 9: Biology uses extensive redundancy to maintain and strengthen connectivity.

Rule 10: Changing biological components and constituents represents a complex undertaking, extending within and across multiple levels of complexity.

Rule 11: Biology optimizes outcomes.

Rule 12: Biology grows in distinct steps, wherein patterns alternate between active growth (dynamic ratios) and no growth (stable ratios).

Rule 13: Biological parts can serve as dominant central organizers, wherein they form connections (ratios) with many other parts.

Rule 14: Biology unfolds into levels of complexity defined by relationships of structure to function.

Rule 15: Reproducibility, which is a universal property of biology, can be detected and verified with ratio-based patterns and equations.

Rule 16: In biology, connectivity generates quantitative patterns.

Rule 17: In biology, a change or disorder involves reordering – by rule - both parts and connections.

Rule 18: In biology, change represents a continuum of complex events involving many parts of an organism.

Rule 19: Biology changes by rule.

Rule 20: In Biology, rules can be identified with equations fitted to regression lines with $R^2 = 1$ or ≈ 1 and as reproducible patterns detected with mathematical markers and connection ratios.

Rule 21: Biochemical homogeneity - a property of biology fundamental to the relationship of structure to function - exists quantitatively as biochemical and morphological densities.

Rule 22: Rate constants define enzyme densities as key elements of biological complexity.

Rule 23: interpreting complex biological changes requires a mathematical collaboration between morphology and biochemistry.

9.4.4 First Principles and Rules (as Equations)

Deriving biology from first principles begins by identifying three such principles:

9.4.4.1 Principle 1: Ratio Rules

Rule 1: Mathematical Marker Rule:

$$\boldsymbol{ax}\boldsymbol{:}\,\boldsymbol{by}\boldsymbol{:}\,\boldsymbol{cz}\,\ldots\boldsymbol{n}$$

Rule 2: Connection Ratio Rule:

$$\boldsymbol{partx}\boldsymbol{:}\,\boldsymbol{party}\boldsymbol{:}\,\boldsymbol{partz}\,\ldots\boldsymbol{n}$$

Rule 3: Ratio Chain Rules:

Morphology

$$\boldsymbol{m_1}\boldsymbol{:}\,\boldsymbol{m_2}\boldsymbol{:}\,\boldsymbol{m_3}\,\ldots\boldsymbol{m_n}$$

Biochemistry

$$\boldsymbol{b_1}\boldsymbol{:}\,\boldsymbol{b_2}\boldsymbol{:}\,\boldsymbol{b_3}\,\ldots\boldsymbol{b_n}$$

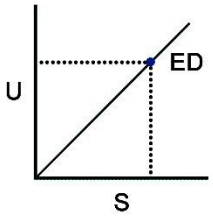
9.4.4.2 Principle 2: Structure-Function Rule

$$\boldsymbol{f(x) = mx; y = mx}$$

Rule 4: Enzyme Density Rule:

$$\boldsymbol{m = y/x}$$

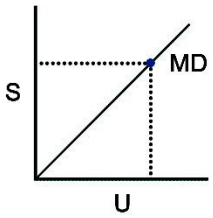
$$\boldsymbol{ED = U/S}$$



Rule 5: Membrane Density Rule:

$$\boldsymbol{m = y/x}$$

$$\boldsymbol{MD = U/S}$$



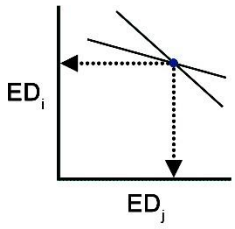
Rule 6: Simultaneous Equations Rule:

Equation 1:

$$[\boldsymbol{S_i} \times \boldsymbol{ED_i}] + [\boldsymbol{S_j} \times \boldsymbol{ED_j} = \boldsymbol{U_{total}}$$

Equation 2:

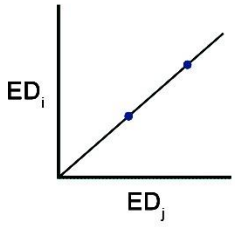
$$[\boldsymbol{S_i} \times \boldsymbol{ED_i}] + [\boldsymbol{S_j} \times \boldsymbol{ED_j} = \boldsymbol{U_{total}}$$



Rule 7: Recipe Detection Rule:

$$\boldsymbol{m = y/x}$$

$$\boldsymbol{m = ED_y/ED_x}$$ **= Ratio**



Rule 8: Recipe Rule:

Control

$$\boldsymbol{aED_{co1}{:}\, bED_{co2}{:}\, cED_{co3} \ldots xED_{con}}$$

Experimental

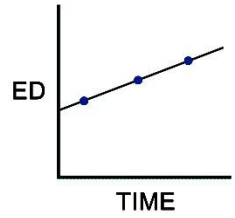
$$\boldsymbol{aED_{ex1}{:}\, bED_{ex2}{:}\, cEDex_3 \ldots xED_{exn}}$$

9.4.4.3 Principle 3: Structure-Function Change Rule

$$\boldsymbol{f}(x) = mx + b$$

Rule 9: Structure-Function Change Rule:

$$ED = (m \times time) + b$$



Rule 10: Rate Equation Rule:

$$[ED] = [ED_0] + kt$$

Rule 11: Relative Change Per Av. Cell Rules:

Total Surface from Total Units

$$S_{total} = MD \times U_{total}$$

Total Units from Total Surface

$$U_{total} = ED \times S_{total}$$

Rule 12: Absolute Change per Organ Rules:

Total Surface from Total Units

$$S_{total} = MD \times U_{total}$$

Total Units from Total Surface

$$U_{total} = ED \times S_{total}$$

By providing access to three first principles and twelve of their derivative rules, biology makes its package of skills available to us along with key information:

- Structure and function are inseparable – one maps mathematically to the other.
- Well-defined relationships of structure to function define complexity, which defines biology.
- Modules embed complexity within complexity - hierarchically and universally.
- Linear models - based on ratios - serve to design and change phenotypes.

Conjectures

- Interpreting molecular events occurring in the genome will require mapping such activities to structural locations in the phenotype.
- The “non-coding” portions of the DNA include many of the command and control codes for the phenotype.

EPILOGUE

What did we learn? Several basic lessons. In attempting to solve biology, we had to address a string of challenging problems – one after the other. Two things quickly became obvious. Most of the challenge was coming from the limitations imposed by our theory structure (reductionism) and by the restrictions built in to the way we report and store our research data (confined to articles published in journals).

Curiously, biology shows no interest whatsoever in reductionism. It runs its business as a complexity, which involves rules, parts, connections, and emergent properties – not just the parts we choose to collect and characterize after taking biology apart. As a complexity, biology operates a big data, rule-based, and highly adaptive enterprise – one that excels at precision, accuracy, reproducibility, adaptability, and success.

By any measure of reason, strategy, or resourcefulness, our one best chance of catching up to biology is to copy it. This involves storing our research data in databases (to encourage data interaction), translating these data into universal formats (e.g., ratios, mathematical markers, connection ratios, equations), and using databases to generate parallel complexities (copies of biology). This puts us on a path, as the primer explains, to emulate biology and thereby become more effective and astute problem solvers.

By playing the rules game, we gain access to biology's first-rate package of skills. We now know how to generate and analyze complex biological patterns, to translate rules into equations, to diagnose and predict, and to verify reproducibility. Moreover, by embracing a first principles approach, biology can begin to teach us the fundamentals of its theory structure.

Why is it to our advantage to play by biology's rules? By redefining reproducibility in line with the way it operates in nature, for example, it reflects reality. In fact, reproducibility defines – in a fundamental way – our approach to discovery by linking local to global such that the rule that applies to an individual also applies to the population. One defines the other. This tells us that data coming from a single paper can be as just as effective as large databases when looking for rules, solutions, and insights.

Another implication exists. When absolute values are not reproducible, ratios or equations derived therefrom most likely are. Moreover, absolute data are not risk free. Since they tend to be the products of "black boxes," they lack the all-important transparency. Although black box data allow us to demonstrate changes statistically, they tell us very little about the underlying mechanisms. So, we opened these boxes, looked inside, and ferreted out the details. Complexity theory provided the means and the mathematical framework.

Such a task became trivial once we understood the relationship between biological variation and reproducibility. This insight allowed us to verify the postulates of biological homogeneity with the structure-function rule ($f(x) = mx$), which defines the relationship of structure to function as a core principle of biological design. Moreover, by simply adding a y intercept (b) to the structure-function rule ($f(x) = mx + b$) , it became possible to unfold the complexity of a biological change as it ripples across the biological hierarchy. In effect, we captured biology's genius for creating complexity with two

wonderfully simple equations. Occam would approve.

The first part of the proof for the biochemical homogeneity rule relied on data coming from three individual animals (biological variation) all of which displayed the same structure-function rule (reproducibility) as an $R^2 = 1$ equation. To pass through the origin – as specified by the rule – the membrane data required the section thickness corrections of Weibel and Paumgartner (1978). Without these corrections, the equations miss the origin and consequently the rule (Bolender, 2018). In turn, predictability and reproducibility combined to produce an effective strategy for validating these structure-function equations.

With access to both the key (the equation encapsulating the postulates of biochemical homogeneity) and the lock (biology's rules expressed as equations), the door between complexity levels 1 and 2 opened wide, thereby allowing us to watch biology as it engineered a series of complex changes.

Two papers (Weibel et al., 1969 and Stäubli et al., 1969) were ideally suited to the task of opening the mysterious black box of cell biology and showing us how multiple levels of complexity appear as hepatocytes responded to the drug phenobarbital. As we watched, biology instigated changes with rule-based equations punctuated throughout by reproducibility.

Since the approach to complexity used for hepatocytes is likely to apply to most if not all cell types, solving biology can proceed smoothly from one cell type to the next. Rediscovering each cell type as a rule-based powerhouse will no doubt lead to a wealth of new and actionable information.

Of course, the story we want biology to tell us is the big one, the one that reveals – mathematically - the relationship of genes to phenotypes to emergent properties. Although we've only come as far as the phenotype, preliminary work suggests that moves into the genotype may require little more than translating or assembling molecular biology databases into parallel complexities.

Complexity quickly becomes the theory of choice when we want to know the implications of making genetic changes, explaining the onset and progression of diseases, designing and testing new drugs, or improving the ability of a species to adapt to new environments. No doubt, such stories - soon to be written - will make for good reading.

We – all of us in the biology business – have been entrusted with two inestimable legacies, the periodic table of elements and the much larger table of genes. Solving biology is just the first step in solving the relationship of one table to the other. This synthesis has already begun and will define our future for many generations to come. Our prospects have never been brighter.

"Study nature, not books."
Louis Agassiz

"… what I tell you three times is true."
Lewis Carroll

GLOSSARY

ABSOLUTE DATA – Data expressed as a volume, surface, length, or number.

ACCURACY – In conformity to truth, a rule, or standard; free from error or defect.

ACID TEST – A conclusive test to establish the value or success of something.

ALGORITHM – A step-by-step sequence of operations designed to perform a specific task.

ALPHANUMERIC – A set (or string) of characters containing letters and numbers.

ANALOGUE – One thing comparable to another.

ARTIFACT – An object made by humans; a distortion produced by an investigative method.

BACK-END – The server side as opposed to the working end (frontend).

LAMBERT-BEER LAW – A method widely used to measure concentrations.

$$log_{10}\frac{I_0}{I} = \epsilon lc,$$

where I_0 is the intensity of the incident light, I the intensity of the emergent light, ε the extinction coefficient, l the length of the light path, and c the concentration.

BIAS – Identifies anything that produces systematic variation in research data; a systematic rather than a random distortion of a statistic.

BIG DATA – Data sets too large to manipulate with traditional methods or tools.

BIOLOGICAL VARIATION – normal differences within and between individuals due to differences in genetic, health, and environmental factors.

BIOLOGICAL CHANGE – It defines a rule-based event that extends across at least six levels of complexity in the phenotype and includes genes, molecules, organelles, cells, organs, and organisms.

BIOCHEMICAL DENSITY (BD) – A complex data type relating a biochemical constituent to a morphological component.

BIOCHEMICAL HOMOGENEITY RULE – Defines the relationship of structure to function in living systems: $f(x) = mx$, where the slope (m) defines and enzyme density (ΔU/ΔS).

BIOCHEMICAL DENSITY RULE – The BD rule defines the relationship of a biochemical constituent (b) to a morphological component (m): $BD = b/m$.

BLACK BOX – A construct involving a system wherein the inner components and logic are unavailable to the observer.

BLUEPRINT – A detailed outline or plan of action; a design.

BUBBLE – Identifies anything that lacks firmness, substance, or permanence; often an illusion or delusion. In biology, they derive from faulty assumptions.

BUTTERFLY – In chaos theory, the butterfly effect exemplifies the dependence of events on initial conditions; a small change can cause a large effect. For example, the turbulence created by a butterfly triggers a storm far away.

CALCULATOR EQUATION – used to extract numerical data from graphs: $y\ value = (((y\_top \times y\_length)/\ (y\_units)) - y\_from\_top) \times (y\_top/((y\_top \times y\_length)/(y\_units\ ))))$

CHAOS THEORY – A branch of mathematics that deals with complex systems. Such systems display

an underlying order, wherein very small events can trigger very complex outcomes.

CLOUD – Provides centralized data storage and retrieval.

CLUSTER ANALYSIS – The grouping of a set of items such that the members of same group (cluster) are closer related to each other than to those in other groups (clusters).

COEFFICIENT OF DETERMINATION – A measure of the goodness of fit between dependent and independent variables in a regression analysis; abbreviated R^2.

COMMUNITYGRAPHPLOT – Identifies related communities (clusters) graphically.

COMPLEX SYSTEMS – Composed of many connected parts. They exhibit properties that emerge from the interaction of their parts, which usually cannot be predicted from the properties of the individual parts.

COMPLEXITY THEORY – Complex behavior emerges from simple rules, producing large networks of interacting parts.

COMPLEXITY LEVELS – Six levels begin the process of unfolding biological complexity into networks of interconnected equations. They include: Level 1 (patterns & changes), Level 2 (biochemical homogeneity), Level 3 (change in organelle), Level 4 (rate of change in organelles), Level 5 (change in cell), and Level 6 (changes in organ).

CONCATENATE – Linking things together in a chain, string, or series.

CONCENTRATION – The amount of a constituent (or component) divided by the total volume of the reference or containing space; expressed per unit volume. Reference spaces can also include surface, length, and number. The definition extends to biochemical densities expressed herein as enzyme and membrane densities.

CONCENTRATION TRAP – Identifies a methodological bias that can occur when attempting to detect a biological change with concentration data. While the volume of the reference compartment (the denominator) remains constant, the number of cells needed to fill the reference volume changes. In effect, both the numerator and denominator can behave as variables.

CONNECTION – Something that connects two or more things. In biology, connections can be defined as ratios derived from the properties of the parts.

CONNECTION PHENOTYPE - Includes a set of parts (data pairs), plotted as a frequency distribution, and fitted to a polynomial regression.

CONNECTION RATIO – A mathematical marker wherein the alpha names in the string are replaced by a single name (“part”); the numerical values of the ratio remain unaltered. This extends the ability of the strings to detect quantitative patterns.

COUNTING MOLECULES – A software package using simulators to explain the pitfalls of interpreting biological changes.

CORRECTED CONCENTRATION EQUATION CORRECTED (CCC) - $CCC_{(t_i)} = C(t_i)\ x\ \frac{W_L(t_i)-W[EHS(t_0)]}{W_L(t_0)-W[EHS(t_0)]}$, where $CCC_{(t_i)}$ is the corrected concentration at time i, $C(t_i)$ the uncorrected concentration at time i, $W_L(t_i)$ the weight of the liver at times i and 0, and $W[EHS(t_0)]$ the weight of the extrahepatocytic space (EHS) at time 0 (t_0). The equation assumes that the EHS remains constant throughout the experiment.

DATA PAIR – A ratio of two numerical values, which may include the names of the parts.

DATA PAIR RULE – A quantitative relationship defined by two parts (A, B) and one connection (X:Y): AX:BY).

DATA STRINGS – A complex data type based on alphanumeric coding; a universal data type used in diagnosis and for detecting patterns and reproducibility.

DATABASE – A structured set of data held in a computer that can be accessed, managed, and updated.

DATA CAGE - A boundary condition imposed by the design of a parallel complexity capable of optimizing outcomes. Such closed systems, for example, were found to be 100% effective for diagnosing disorders of the brain. Moreover, a data set contained within such a cage becomes predictive when allowed to interact with outside data.

DATA-DRIVEN – Progress propelled by data, rather than by methods.

DECIMAL REPERTOIRE EQUATION – The values of a repertoire equation fitted to decimal steps.

DENSITY – A term used in stereology to describe a concentration.

DESCRIPTIVE BIOLOGY – A qualitative approach to biology.

DESIGN-BASED SAMPLING – Sampling independent of size, shape, orientation, and distribution; sampling bias is minimized. Every part of the structure has the same chance of being sampled.

DIAGNOSE – Identify a normal or abnormal state usually by examining symptoms.

DIAGNOSIS DATABASES: Usually based on triplets, these databases use unique or duplicate markers. They are expected to play a key role in explaining how the genome orchestrates phenotypic changes.

DESIGN CODES – Include ratios formed by dividing experimental by control values. They identify patterns of change.

DESIGN CODE EQUATIONS – Include the use of regression analysis to look for linear patterns in biological data. Such patterns suggest the presence of underlying rules.

DISEASE – A disorder of structure or function, often producing specific signs and symptoms.

DISECTOR – A design-based method of stereology that uses an unbiased sampling frame to estimate the numerical density (N/V) of particles.

DISORDER – A malady or dysfunction; a state of confusion.

DISRUPTION – To break apart or alter, thereby preventing the existence of a normal.

DISTORTED – Not representing the facts or reality; misrepresenting; false.

DUPLICATE – One of two or more identical things.

EMERGENT PROPERTY – Connected parts display new properties equal to more than those of the individual parts; the whole is greater than the sum of the parts; properties irreducible to the constituent parts.

EMPIRICAL – Identifies outcomes based on testing or experience rather than on theory.

ENCAPSULATE – Enclose; contain.

ENTERPRISE – A project, one that is difficult or requires considerable effort.

ENTERPRISE BIOLOGY SOFTWARE PROJECT (EBSP) – A project designed specifically to speed learning and discovery in the life sciences (enterprisebiology.com).

ENZYME DENSITY (ED) – A ratio (U/S) defining the relationship of an enzyme activity (U) to a morphological reference (usually a surface area (S)). It provides the amount of activity associated with 1 m^2 of membrane surface area.

ENZYME DENSITY RULE: An example of the biochemical homogeneity rule, it defines a quantitative relationship that exists between units of marker enzyme activity (U) and 1 m^2 of membrane surface area (S): $ED = U/S$.

EQUATION – An expression stating that two things are equal.

ERROR – A deviation from accuracy or correctness.

EXPERIMENT – A scientific procedure designed to discover, test a hypothesis, or demonstrate a known fact.

EXPONENT – The power to which a given number or expression is to be raised.

FALSE NEGATIVE – Indicates mistakenly that something tested for is absent when it is present.

FALSE POSITIVE – Indicates mistakenly that something tested for is present when it is not.

FILTER – A device designed to remove specific components.

FIRST PRINCIPLE: A first principle can be a law upon which others are founded or from which others are derived. It is a general truth, comprehending many subordinate truths, but not deductible from others.

FOLD – To place together and entwine; to blend components; to bring from extended to compact.

FRACTIONATOR – A design-based method of stereology used for estimating particle counts; based on the disector; a systematic random sampling method.

FRONT-END – User interface; the part of a software program with which the user interacts.

GAME – structured playing; a form of sport or play; played according to rules and decided by skill, strength, or luck.

GENERALIZATION – A general statement, law, principle, or proposition.

GENOTYPE – Genetic constitution of an individual.

GLOBAL – Involving all of something.

GOLD STANDARD – The example by which others are judged or measured.

HIERARCHY – A series of ordered groupings.

HOMOGENATE – Biological tissue ground up and mixed; mechanically disrupted.

IBVD – Internet Brain Volume Database; MRI data expressed as volumes.

INTACT TISSUE – Undamaged; unaltered.

JOIN – Cause to become joined or linked, as in database tables.

LADDER EQUATION – An exponential equation summarizing a set of rung (power) equations.

MASHUP – A method for discovering new types of information by combining data from one or more sources.

MATHEMATICA – A computational software program; Wolfram Research, Champaign, Ill.

MATHEMATICAL CORE – Used herein to identify the quantitative rules to which biology adheres.

MATHEMATICAL MAPPING – An element of a given set associated with an element of another set.

MATHEMATICAL MARKER – An alphanumeric string designed to captures units of complexity specific to a given phenotypic state.

MEMBRANE DENSITY (MD) – A ratio (S/U) defining the amount of membrane surface area (S) supporting one unit (U) of enzyme activity.

MEMBRANE DENSITY RULE - An example of the biochemical homogeneity rule, it defines a quantitative relationship that exists between 1 m^2 of membrane surface area (S) and units of marker enzyme activity (U): $MD = S/U$.

METHODS-DRIVEN – An activity compelled by methods.

NESTED COMPLEXITY – Complexity embedded in complexity. Unfolding and refolding nested complexity represents a major undertaking of complexity theory. The process consists of translating data sets into mathematical markers, storing them in a universal biology database, and applying filtering algorithms.

NETWORK – A system of interconnected parts.

OBJECTIVE – Not influenced by personal feelings or opinions; identified with quantitative approaches.

OPTICAL DENSITY – A measure of the extent to which a substance transmits light or other electromagnetic radiation.

ORGANISM CODES – Identify patterns of connectivity in each paper graphically.

PARALLEL COMPLEXITY – A collection of mathematical markers serving as a proxy for biology; a construct designed with a specific goal in mind (e.g., diagnosis). It serves as a copy of biology.

PATTERN – A repeated design; an arrangement or sequence; things arranged by rule.

PERMUTATION – The way in which a set of numbers or things can be ordered.

PHENOTYPE – The physical appearance of an organism. It represents the downstream expression of a genetic code, represented by the observable characteristics of an organism such as morphology, biochemistry, physiology, or behavior.

PHENOTYPIC DATA – Genetic expression detected herein, for example, as mathematical markers, connection rations, rules, and equations.

PLAYING FIELD – A database platform for playing complexity games with properties specified according to the game's rules; a field designed to solve a specific problem.

POLYNOMIAL EQUATION – An algebraic equation wherein one or both sides take the form of a polynomial.

POSTMORTEM – After death.

POWERBUILDER – An integrated development environment distributed by Sybase, Inc. (Emeryville, CA).

PRECISION – The extent to which a given set of estimates of the same sample agree with the mean value.

PREDICT – To tell in advance, using, for example, inference or special information.

PROPORTION – A relationship among parts with respect to their comparative quantity.

PROXY – A substitute for another.

QUADRUPLET MARKER – A mathematical marker consisting of four alpha and four numeric components; expressed as a numerical ratio.

QUADRUPLET RULE – A mathematical marker (expressed as decimal values): AXBYCZDQ.

QUALITATIVE – Distinctions based on qualities.

QUANTITATIVE – Expressible as a quantity; susceptible to measurement.

QUERY BY EXAMPLE (QBE) – Query by example; a database query based on the items selected.

$R^2 = 1$ EQUATIONS – Used to translate complex data and rules into equations.

RATE EQUATIONS – Defines the speed at which a concentration is changing.

RATE EQUATION RULE – A chemistry rule ($[A] = [A]_0 \pm kt$) adapted to biology ($[ED] = [ED]_0 \pm kt$), wherein enzyme densities become the concentrations.

RATIO – Relative magnitudes of two or more quantities.

RATIO CHAIN RULE – A rule that applies to both morphology ($m_1 : m_2 : m_3, \ldots, m_n$) and biochemistry ($b_1 : b_2 : b_3, \ldots, b_n$).

RECIPE DETECTION RULE – A rule that can be applied during the transitional state of a biological change to detect the appearance of a new recipe – expressed as an alphanumeric string of parts and connections.

RECIPE RULE: An alphanumeric string of enzyme densities (expressed as ratios) identifies the phenotypic code defined by gene expression. Such codes map back to specific genes and to yet unidentified portions of the DNA.

REDUCTIONIST THEORY – Assumes that complex systems can be understood in terms of their individual parts.

REGRESSION EQUATION – The relationship between values X and Y from which the most probable value of Y can be predicted from X.

RELATIONSHIPS OF STRUCTURE TO FUNCTION – Play a key role in defining biological complexity. They express rules that can be captured with equations.

RELATIVE AND ABSOLUTE CHANGE RULES: These structure-function rules monitor changes indexed to total cell and organ changes.

REPERTOIRE EQUATION – Defines the quantitative relationship of values X to Y, wherein both the slope and the R^2 of a power curve approach one. A repertoire equation identifies a relationship defined by data published in the biology literature.

REPRESENTATIVE SAMPLE – A population that accurately reflects the members of the entire population.

REPRODUCIBILITY – to produce again; to duplicate.

REPRODUCIBILITY (STANDARD DEFINITION) – An ability to repeat the results of an experiment either by the same researcher or by an independent one.

REPRODUCIBILITY (UPDATED DEFINITION) - An ability to duplicate a biological complexity with little or no biological variation.

ROSETTE – a visual representation of complexity, consisting of a central point (connection ratio) connected by lines (forming a petal) to a mathematical marker. It serves as a Level 1 test of reproducibility for large data sets.

RULE-BASED – A production system based on rules for storing, manipulating, and interpreting information in a useful way.

RUNG EQUATION – Data fitted to a power curve displaying an R^2 approaching one.

SCALE – To change a process to allow for greater or lesser quantities; increase or decrease proportionately.

SCATTERPLOT – A haphazard distribution of data; bivariate data expressed as a set of scattered points.

SCIENCE – Extends knowledge of principles and causes.

SEMIQUANTITATIVE – Partially quantitative and partially qualitative; approximating a quantitative value.

SHRINKAGE – The amount by which something decreased in amount, often a volume; a reduction in value.

SIGNIFICANT DIFFERENCE – A measure of the likelihood of drawing a false conclusion in a statistical test; too closely correlated to be attributed to chance alone; implies a measurable difference between two groups wherein the probability of obtaining that difference by chance is very small (e.g., <0.05; <5%).

SIMPLE – Having few parts; not complex or complicated.

SIMULTANEOUS EQUATIONS – Includes a method for finding solutions based on the intersection of two linear equations. It offers a way of finding unknown enzyme densities.

SIMULATION – Act of imitating the behavior of some situation or process; to create a representation or model.

STANDARDIZE – To conform to a representative process or norm.

STANDARD – A basis for comparison.

STEADY STATE RULE – Identifies a biochemical homogeneity expressed as an enzyme or membrane density: $f(x) = mx$, where the slope (m) is the density.

STEREOLOGY – A collection of mathematical methods for estimating 2D and 3D structures quantitatively from lower dimensions.

STEREOLOGY LITERATURE DATABASE (SLD) – published research data stored in a relational database.

STOICHIOMETRY – Relationships existing as a ratio of small integers.

STRUCTURE-FUNCTION RULE (BIOCEMICAL HOMOGENEITY RULE) – Defines a quantitative relationship between structure and function ($f(x) = mx$), where m is the slope of the curve passing through the origin. It represents a principle fundamental to the design and operation of biology. It defines enzyme and membrane densities.

STRUCTURE-FUNCTION CHANGE RULE – Identifies a biological change as a relationship of structure to function: $f(x) = mx + b$. It is the product of several individual biochemical homogeneity (i.e., enzyme density) rules: $f(x) = mx$.

SUBJECTIVE – Coming more from the observer than from observations.

SWELLING – An increase in volume.

SYMPTOM – Indicates the presence of something, especially something undesirable.

SYNDROME – A pattern of symptoms indicative of a disease.

SYSTEM – A group of independent, but interrelated elements producing a unified whole.

TEMPLATE – A guide for making other objects.

THEORETICAL – Concerned with theories, rather than practical applications.

THEORY – A well-substantiated explanation of some aspect of the natural world.

TRADITIONAL – Time-honored orthodox doctrines; widely accepted.

TRANSITIONAL STATE (Change) – Defined by the structure-function rules, it identifies a period of time when a biological change occurs.

TRIPLET MARKER - A mathematical marker consisting of three alpha and three numeric components; expressed as a numerical ratio.

TRIPLET RULE - A mathematical marker (expressed as decimal values): AXBYCZ.

TRIPLET DATABASES – Populated with mathematical markers (AXBYCZ), these databases serve to identify patterns in large data sets.

UNBIASED – When bias equals zero; lack of systematic error.

UNBIASED DATA – When bias equals zero for the method of sampling and the material sampled.

UNBIASED SAMPLING – A method designed to remove bias from the sampling procedure; design-based sampling.

UNFOLD – Open out; to reveal or display; lay open to view.

UNIQUE – One of a kind.

UNIT – A standard of measurement; appended to a value.

UNIVERSAL BIOLOGY DATABASE – Contains biological data expressed as ratios; a unified data set derived from the biology literature.

UNIVERSAL STANDARDS – Identifies a best practices approach to biological research. They include universal biology databases, universal connectors, universal references for morphology and biochemistry, and universal gold standards.

UNSTABLE – Lacking stability; affords no assurance; subject to change, variable, unpredictable, ambiguous.

VALENCE – An ability of a given part to connect to the same part in different ratios.

VALUE – A numerical quantity.

VARIABLE – A quantity that can assume different values.

WEIBEL-PAUMGARTNER CORRECTION – Corrects stereological estimates for the biases produced by section thickness. Stereological estimates assume that the sections used to collect data have no thickness.

WIN-WIN – An outcome beneficial to all parties involved.

ZEROTH-ORDER REACTION: Defines a linear equation ($[A] = [A]_0 \pm kt$), where [A] is the concentration a given time, $[A]_0$ the concentration at time 0, k the rate constant, and t the given time.

BIBLIOGRAPHY

Aase S., Roland M., and BR Olsen. (1976) Ultrastructure of parietal cells before and after proximal gastric vagotomy in duodenal ulcer patients. Scand J Gastroenterol 11: 55-64.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P. Molecular Biology of the Cell 6th Edition (2014) Garland Science, New York; 1464 pages.

Amar-Costesec A., Beaufay H., Wibo M., Thinès-Sempoux D., Feytmans E., Robbi M., Berthet J. (1974) Analytical study of microsomes and isolated subcellular membranes from rat liver. II. Preparation and composition of the microsomal fraction. J Cell Biol. 61:201-212.

Baker, M. (2016) Is There A Reproducibility Crisis? Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a; Nature's Questionnaire (2016) : http://www.nature.com/polopoly_fs /7.36741%21/ file/Reproduciblility%20Questionnaire.doc

Begley, C.G., and Ioannidis, J.P. (2015) Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. Circ Res 116: 116126.doi: 10.1161/CIRCRESAHA. 114.303819 PMID: 25552691

Bertram, J. F., P. D. Sampson, Bolender, R.P. (1986) Influence of tissue composition on the final volume of rat liver blocks prepared for electron microscopy. J. Electron Microsc. Tech. 4: 303-314.

Blouin A., Bolender R.P., Weibel E.R. (1977) Distribution of organelles and membranes between hepatocytes and nonhepatocytes in the rat liver parenchyma. A stereological study. J Cell Biol. 72(2):441–455.

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

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

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

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

Bolender, R. P. (2010) Enterprise Biology Software XI. Research 2010 In: Enterprise

Biology Software, Version 10.0 © 2010 Robert P. Bolender

Bolender, R. P. (2011) Enterprise Biology Software XII. Research 2011 In: Enterprise Biology Software, Version 11.0 © 2011 Robert P. Bolender

Bolender, R. P. (2012) Enterprise Biology Software XIII. Research 2012 In: Enterprise Biology Software, Version 12.0 © 2012 Robert P. Bolender

Bolender, R. P. (2013) Enterprise Biology Software XIV. Research 2013 In: Enterprise Biology Software, Version 13.0 © 2013 Robert P. Bolender

Bolender, R. P. (2014) Enterprise Biology Software XV. Research 2014 In: Enterprise Biology Software, Version 14.0 © 2014 Robert P. Bolender

Bolender, R. P. (2015) Enterprise Biology Software XVI. Research 2015 In: Enterprise Biology Software, Version 15.0 © 2015 Robert P. Bolender

Bolender, R. P. (2016a) Enterprise Biology Software XVII. Research 2016 In: Enterprise Biology Software, Version 16.0 © 2016 Robert P. Bolender

Bolender, R.P. (2016b) Playing the Complexity Game with Biology. © 2016 Enterprise Biology Software Project, PO Box 292 Medina, WA 98039-0292

Bolender, R. P. (2017) Enterprise Biology Software XVIII. Research 2017 In: Enterprise Biology Software, Version 17.0 © 2017 Robert P. Bolender

Bolender, R. P. (2018) Enterprise Biology Software XIX. Research 2018 In: Enterprise Biology Software, Version 18.0 © 2018 Robert P. Bolender

Bolender, R. P. and Charleston J. S. (1993) Software for counting cells and estimating structural volumes with the optical volume fractionator. Microsc. Res. Tech. 25: 314-324.

Bolender, R. P. and Bluhm J.M. (1992) Database literature review: A new tool for experimental biology. Mathl. Comput. Modelling, 16:11-35.

Bolender R. P., Paumgartner D., Muellener D., Losa G., Weibel E. R. (1978) Integrated stereological and biochemical studies on hepatocytic membranes. I. Membrane recoveries in subcellular fractions. J Cell Biol. 1978 May 1; 77(2): 565–583. (Free download)

Bolender R. P., Paumgartner D., Muellener D., Losa G., Weibel E. R. (1980) Integrated stereological and biochemical studies on hepatocytic membranes. IV. Heterogeneous distribution of marker enzymes on endoplasmic reticulum membranes in fractions. J Cell Biol. Jun;85(3):577-86.

Bolender R. P. and Weibel E. R. (1973) A morphometric study of the removal of phenobarbital-induced membranes from hepatocytes after cessation of treatment. J Cell Biol. 56(3):746-761.

Bolender, R. P., Hyde, D. M., DeHoff R.T. (1993) Lung morphometry: a new generation of tools and experiments for organ, tissue, cell, and molecular biology. Am. J. Physiol. 265 (Lung Cell. Mol.Physiol. 9): L521L548.

Cecil K.M., Brubaker C.J., Adler C.M., Dietrich K.N., Altaye M., Egelhoff J.C. , Wessel S., Elangovan I., Hornung R., Jarvis, K., Lanphear B.P. (2008) Decreased Brain Volume in Adults with Childhood Lead Exposure. PLoS Med 5(5): e112. doi:10.1371/journal.pmed.0050112

Collins, F.S. and Tabak L.A. (2014) NIH plans to enhance reproducibility. Nature 505:612–613.PMID: 24482835

Colquhoun D. (2014) An investigation of the false discovery rate and the misinterpretation of p-values. R. Soc. Open sci. 1: 140216. http://dx.doi.org/10.1098/rsos.140216

Cruz-Orive, L. M. and Weibel E.R. (1990) Recent stereological methods for cell biology: a brief survey. Am. J. Physiol. 258 (Lung Cell. Mol. Physiol. 2) L148-L156.

de Duve, C. Principles of tissue fractionation. J Theor Biol. (1964) Jan;6(1):33–59.

de Duve, C. Exploring cells with a centrifuge. (1974) Nobel Lecture, December 12, 1974

Engber, D. (2016) Cancer Research Is Broken - There's a replication crisis in biomedicine—and no one even knows how deep it runs. Future Tense, April 19, 2016 FROM SLATE, NEW AMERICA, AND ASU

Freedman, L.P., Cockburn, I.M., Simcoe, T.S. (2015) The Economics of Reproducibility in Preclinical Research. PLOS Biology|DOI:10.1371/ journal.pbio.1002165

Goldstein, J.M., Goodman, J.M., Seidman, L.J., Kennedy, D.N., Makris, N., Lee, H., Tourville, J., Caviness, Farcone, S.V., Tsuang M.T. (1999) Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. Arch Gen Psychiatry 56, 537-547.

Guido K.W. Frank G.K.W., Shott, M.E., Hagman J.O., Yang T.T. (2013). Localized Brain Volume and White Matter Integrity Alterations in Adolescent Anorexia Nervosa. Journal of the American Academy of Child & Adolescent Psychiatry 52: Issue 10, 1066– 1075.e5.

Gundersen H. J. G., Bagger P., Bendtsen T. F., Evans S. M., Korbo L., Marcussen N., Moller A., Nielsen K., Nyengaard J. R., Pakkenberg B., Sorensen A., Vesterby, West M.J. (1988) The new stereological tools: disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis. Acta Pathol. Microbiol. Immunol. Scand. 96: 857-881.

Harding A.J., Wong A., Svoboda M., Kril J.J., Halliday G.M. (1997) Chronic alcohol consumption does not cause hippocampal neuron loss in humans. Hippocampus 7:78-87.

Herting M.M., Gautam P., Spielberg J.M., Kan E, Dahl R.E., Sowell E.R. (2014) The role of testosterone and estradiol in brain volume changes across adolescence: A longitudinal structural MRI study. Human Brain MappingVolume 35, Issue 11, 5633–5645.

Hinchliffe S.A., Sargent P.H., Howard C.V., Chan Y.F., van-Velzen D. (1991) Human intrauterine renal growth expressed in absolute number of glomeruli assessed by the disector method and Cavalieri principle. Lab Invest 64: 777-784.

Ioannidis, J.P.A. (2005) Why most published research findings are false. PLoS Med 2: e124.

Ioannidis J.P. and Panagiotou O.A. (2011) Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA. 2011; 305 (21): 2200-10.

Kauffman, S. At Home in the Universe. (1995) (Oxford University Press), New York.

Kennedy D. N., Hodge S. M., Gao Y., Frazier J. A., Haselgrove C. (2012) The internet brain volume database: a public resource for storage and retrieval of volumetric data. Neuroinformatics. Apr; 10(2):129-40.

Khan O., Bao F., Shah M., Caon C., Tselis A., Bailey R., Silverman B., Zak I., Resnick S.M., Goldszal A.F., Davatzikos C., Golski S., Kraut M.A., Metter E.J., Bryan R.N., Zonderman A.B. (2012) Effect of disease modifying therapies on brain volume in relapsing remitting multiple sclerosis: results of a five-year brain MRI study. J Neurol Sci. 312(1-2):7-12. doi: 10.1016/j.jns.2011.08.034. Epub 2011 Sep 13.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A., Martin, K. C. (2016) Molecular Cell Biology 8th Edition, W. H. Freeman, New York.; 1170 pages.

Losa G. A., Weibel E. R., Bolender R. P. (1978) Integrated stereological and biochemical studies on hepatocytic membranes. III. Relative surface of endoplasmic reticulum membranes in microsomal fractions estimated on freeze-fracture preparations. J Cell Biol. 1978 78: 289-308.

Kauffman, S. At Home in the Universe. (1995) (Oxford University Press), New York.

Malendowicz K. (1986) A correlated stereological and functional studies on the long-term effects of ACTH on rat adrenal cortex. Folia Histochem Cytobiol 24: 203-211.

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

Murphy, K. (2016) Should All Research Papers Be Free? The New York Times, March 12, 2016

National Research Council, National Academy Press, Washington, D.C.

Marcos, R., Monteiro, R.A.F., Rocha, E. (2006) Design-based stereological estimation of hepatocytic number, by combining the smooth optical fractionator and immunocytochemistry with anti-carcinoembryonic antigen polyclonal antibodies. Liver International 26: 116-124.

NIH Research Funding Trends: http://faseb.org/Science-Policy-and-Advocacy/Federal-Funding-Data/NIH-Research-Funding-Trends.aspx

Nikicicz H., Kasprzak A., Malendowicz L.K. (1984) Sex differences in adrenocortical structure and function. XIII. Stereologic studies on adrenal cortex of maturing male and female hamsters. Cell Tissue Res 235: 459-462.

Orrenius, S., and Ericsson, L.E. (1965) Phenobarbital induced synthesis of the microsomal drug-metabolizing enzyme system and its relationship to the proliferation of endoplasmic membranes. A Morphological and biochemical study. J Cell Biol. 25:627-639.

Pellizzari E., Lohr K., Blatecky A., Creel, D. (2017) Reproducibility: A Primer on Semantics and Implications for Research. RTI International (Research Triangle Park, NC).

Poline J-B, Breeze JL, Ghosh S, G Krzysztof OH, Halchenko YO, Hanke M, Haselgrove C, Helmer KG,Keator DB, Marcus DS, Poldrack RA, Schwartz Y, Ashburner J, Kennedy D.N. (2012) Data sharing in neuroimaging research. Frontiers in Neuroinformatics. 2012;6:9. doi:10.3389/ fninf.2012.00009.

Roth, K.A. and Cox, A.E. (2015) Science Isn't Science If It Isn't Reproducible. Am J Pathol 185:2–3.doi:10. 1016/j.ajpath.2014.11.001 PMID:25529794

Schwanhaüsser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M. (2011) Global quantification of mammalian gene expression control. Nature 473: 337-342.

Seecharan, D.J., Kulkarni, A.L., Lu, L., Rosen, G.D., Williams R.W. (2003) Genetic control of interconnected neuronal populations in the mouse primary visual system. Neurosci 23: 11178-88.

Stäubli, W., Hess, R., Weibel, E.R. (1969) Correlated morphometric and biochemical studies on the liver cell. II Effects of phenobarbital on rat hepatocytes. J Cell Biol. 42:92-112.

Strassburger T.L., Lee H.C., Daly E.M., Szczepanik J., Krasuski J.S., Mentis M.J., Salerno J.A., DeCarli C., Schapiro M.B., Alexander G.E. (1997) Interactive effects of age and

hypertension on volumes of brain structures. Stroke. 28(7):1410-7.

Tiehuis A.M., van der Graaf Y., Visseren F.L., Vincken K.L., Biessels G.J., Appelman A.P., Kappelle L.J., Mali W.P. (2008) Diabetes increases atrophy and vascular lesions on brain MRI in patients with symptomatic arterial disease. Stroke. 39(5):1600-3. doi: 10.1161/STROKEAHA.107.506089. Epub 2008 Mar 27.

Van Regenmortel, M.H.V. (2004) Reductionism and complexity in molecular biology. EMBO 5(11): 1016–1020. doi:10.1038/sj.embor.7400284 PMCID: PMC1299179

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

Weibel, E.R. (1979) Stereological Methods, Vol. 1. Practical Methods for Biological Morphometry. (Academic Press), London.

Weibel E. R., Paumgartner D. (1978) Integrated stereological and biochemical studies on hepatocytic membranes. II. Correction of section thickness effect on volume and surface density estimates. J Cell Biol. 1978 77: 584-597.

Weibel, E.R., Stäubli, W., Gnägi, H. R., Hess, R. (1969) Correlated morphometric and biochemical studies on the liver cell. I Morphometric model, stereologic methods, and normal morphometric data for rat livers. J Cell Biol. 42:68-91.

West, M.J. (2012) Basic Stereology for Biologists and Neuroscientists, Cold Spring Harbor Laboratory Press, New York.

Zhu, L., Kim, S.-J., Hara, M., Aono, M. (2018) Remarkable problem-solving ability of unicellular amoeboid organism and its mechanism. R. Soc. open sci. 5: 180396.

INDEX

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

Z