



PLAYING THE COMPLEXITY GAME WITH BIOLOGY

Robert P Bolender

Playing the Complexity Game with Biology

Robert P Bolender

Copyright © 2016 by Robert P. Bolender
All rights reserved. This book or any portion thereof
may not be reproduced or used in any manner whatsoever
without the express written permission of the publisher
except for the use of brief quotations in a book review.

Printed in the United States of America

First Printing, 2016

ISBN 978-0-9971543-0-6

Enterprise Biology Software Project
PO Box 292
Medina, WA 98039-0292

www.enterprisebiology.com

Contents

Preface	8
Acknowledgments.....	11
Introduction	12
The Problem	12
Theory Structure	12
Games	14
Background	14
Challenges	16
Story	16
Perceptions	17
Bigger Picture	18
Language	18
Chapter 1.....	19
Game 1 – Reconnecting Data.....	19
1-1 Relational Database	19
1-2 Entering Data	20
1-3 Data Catalogues	21
1-4 Reductionist Theory.....	22
1-5 Playing Fields.....	22
1-6 Concentrations.....	22
1-7 Absolute Values	24
1-8 Design Code Equations	26
1-9 Summary of Chapter 1	27
Chapter 2.....	28
Game 2 – Finding the Rules	28
2-1 A New Role Model	28
2-2 Data Pairs (Ratios).....	28
2-3 Universal Biology Database	29
2-4 Repertoire Equations	31
2-5 Ladder Equations	32
2-6 Rung Equations	33

2-7 Simulators	34
2-8 Reverse Engineering	35
2-9 Decimal Ratios and Decimal Repertoire Equations	36
2-10 Extracting Hidden Information	36
2-11 Growth Kinetics.....	38
2-12 Biological Blueprint.....	38
2-13 Connection Phenotypes.....	41
2-14 Data Triplets.....	44
2-15 Organism Codes	45
2-16 Fibonacci Numbers	47
2-17 Summary of Chapter 2	48
Chapter 3.....	49
Game 3 – Creating a Parallel Complexity.....	49
3-1 Mathematical Mapping	49
3-2 Diagnostic Patterns.....	51
3-3 Game Changer	52
3-4 Mathematical Markers	52
3-5 Diagnosing Disorders of the Brain (Shared Markers)	53
3-6 Generalizing Disorders.....	56
3-7 Playing the Disorder Game	57
3-8 Summary of Chapter 3	59
Chapter 4.....	60
Game 4 – Reconciling Differences	60
4-1 Diagnosing Disorders Post-mortem.....	60
4-2 Global Patterns in Normal Brains	61
4-3 Disrupted Global Patterns in the Brain.....	62
4-4 Artificial Complexity.....	63
4-5 Reality Check.....	64
4-6 Corrections for Post-mortem Data	64
4-7 Summary of Chapter 4	67
Chapter 5.....	68
Game 5 - Diagnosing Disorders of the Brain	68
5-1 Technology Shift	68

5-2 Filtering Mathematical Markers	68
5-3 Test 1: Quadruplets (Shared Markers)	70
5-4 Test 2: Triplets (Shared Markers)	72
5-5 Test 3: Triplets (Unique Markers)	73
5-6 Test 4: Quadruplets (Unique Markers)	74
5-7 Test 5: Quadruplets (Unique Markers)	75
5-8 Test 6: Quadruplets (Unique Markers)	76
5-9 Test 7: Triplets (Unique Markers)	77
5-10 Summary of Chapter 5	78
Chapter 6.....	80
Game 6 – The Disease Process.....	80
6-1 Unfolding the Complexity of Disease.....	80
6-2 Generalizing Disorders with Modular Markers.....	80
6-3 Finding Communities of Disorders	81
6-4 Sharing Markers.....	84
6-5 Sharing Markers and Symptoms.....	85
6-6 Identifying the Prime Movers	87
6-7 Summary of Chapter 6	91
Chapter 7.....	92
Caveats.....	92
7-1 Change	92
7-2 Stereology	92
7-3 Data Points.....	92
7-4 Data Equivalency.....	92
7-5 Volume Independent Methods.....	93
7-6 Reproducibility.....	93
7-7 Biological Variation	93
7-8 Experiments	95
7-9 Modifying the Human Genome	95
7-10 Published Research Findings	95
7-11 Biology as a Science	96
7-12 Summary of Chapter 7	96
Chapter 8.....	97

Theory of Biological Complexity	97
8.1 Introduction	97
8.2 A First Principles Approach	97
8.3 Theory of Biological Complexity	98
8.4 Theory Structure	98
8.4.1 Goals.....	98
8.4.2 Data Requirements	98
8.4.3 Basic Principles and Definitions	99
8.4.4 Derivatives	99
8.4.5 First Principles (Rules)	100
Chapter 9.....	101
Recommendations	101
9.1 Background	101
9.2 Strategy	101
9.3 Issues.....	101
9.4 Current Reality	101
9.4 Recommendations	102
9.4.1 Theory Structure	102
9.4.2 Technology	102
9.4.3 Publication	102
9.4.4 Data Management	102
9.4.5 Data Interpretation	102
9.4.6 Support.....	102
Epilogue.....	103
Glossary.....	106
Working Definitions	106
Bibliography	109
Index.....	112

Preface

This is an adventure story, intent on going to curious places and engaging problems difficult enough to instigate new approaches to problem solving. To keep things interesting, we will deliberately increase the risk of our adventure by getting ourselves into seemingly impossible situations on the assumption that the deeper the trouble the better the story. We will do this by creating chaos as we wander fortuitously from one problem to the next. Much to our surprise, this seemingly aimless approach will serve us well in that it will teach us that the main part of our job is to figure out that solutions to some of our most pressing problems already exist. To enliven the story further, we will cast the principal player – biology – as both hero and antihero by juxtaposing it, as it exists to how we think it exists.

We already know what happens when we take biology apart, but we have absolutely no idea what to expect when we put it back together. Since this is exactly what we are about to do, we find ourselves face to face with one of the most intimidating problems imaginable – biological complexity.

The first thing to know about complexity is that it comes with its own set of rules. It considers many of our current rules as bubbles, well ripened and ready to burst. Success, we will discover, often requires little more than simply changing our perspective from upside down to right side up. Acceptable can become unacceptable and unacceptable acceptable.

Since busting bubbles can have serious consequences, we must proceed prudently. To be fair, we agree at the outset to fix whatever we break. As we work our way through several bubbles, the narrative will accumulate a body of evidence suggesting that our current approach to complex problem solving in biology is sadly amiss – largely because it relies heavily on a theory structure bound tightly to reductionism.

Here is the problem. We have a science – called biology – that lacks a mathematical foundation and can produce data so corrupted by bias and biological variation that the original information often becomes unrecognizable. To make matters worse, we assume that we can study biology by reducing its complexity to a simplicity, characterize its parts in isolation, and then use the resulting information to explain biology as it normally exists. We dig the hole even deeper by assuming that our methods allow us to detect biological changes, when often the best they can do is detect significant differences between heavily biased data sets. Consequently, the data we publish all too often stand little chance of representing biology, as it is.

Now, we come to the more challenging part of our story. Biology exists as a mathematical powerhouse running systems so complex that they defy even our imagination. In short, biology uses rules and algorithms to produce and maintain a complexity that we call a phenotype (Figure P.1).

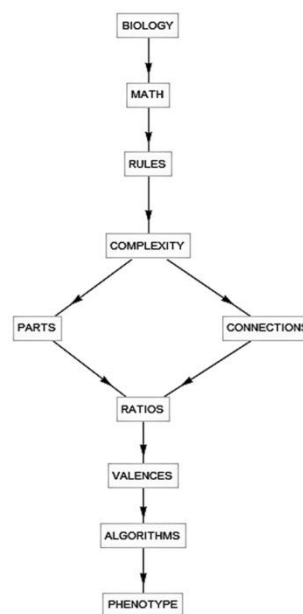


Figure P.1 Biology follows a rule-based approach for producing, maintaining, and adapting phenotypes. Our adventure becomes one of finding a mathematical route

from phenotype to genotype, using published data as our compass.

For biology, a phenotype is an optimized version of an intelligent, battle tested, complex self-adaptive system. It represents nature at its best. For us, the phenotype remains largely inaccessible because of our collective indifference to biology as a complexity. Given the information in Figure P.1, however, we now have a road map to this phenotype with all the arrows pointing in the right directions. To embrace complexity and reinvent biology as a quantitative science, all we have to do is duplicate Figure P.1 using data from the biomedical literature (Figure P.2) – provided we can resolve the thorny issue of data access.

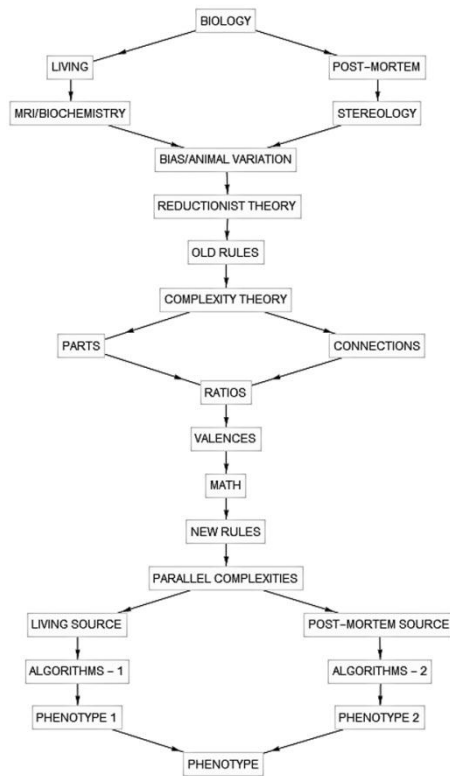


Figure P.2 Reading a phenotype mathematically involves managing the destabilizing effects of our experimental methods and learning the rules of complexity from biology.

The first thing we notice after comparing these two figures is that our job looks even harder than the one belonging to biology. By taking

biology apart to study it, we unwittingly add multiple levels of complexity to an already complex biology. Before we can access the phenotype, we have to delete the extraneous complexities and then figure out how to read biology mathematically.

As the story unfolds, we will eventually discover that it takes a complexity to solve a complexity. Since we – as investigators - have little or no practical or theoretical experience in dealing with biological complexity, we will have to come up with a new theory structure for biology, one that will guide the way. In time, we will identify a parallel complexity as a major problem solver because it effectively recruits biology to do most of the heavy lifting for us. A theory structure capable of producing these parallel complexities gives us the advantage of being able to interact with biology mathematically. If we have a problem and can set it up correctly, biology always seems to have a solution waiting for us.

Our foray into complexity seems well timed in that the biology community currently finds itself under attack from our statistical colleagues. In a scathing article, Ioniades (2012) suggests that as few as 20% of our published papers may be correct, whereas Colquhoun (2014) puts it at 30%. These are serious people making serious allegations. Moreover, a recent editorial in the Journal of Basic and Applied Social Psychology openly rejects the prevailing view that a significant difference in biology can be set at the 95% ($P \leq 0.05$) level. In fact, the journal no longer accepts papers unless they reach a significance level of 99% ($P \leq 0.01$). If we applied such a rigorous standard to our biomedical literature, many of our published research papers would effectively disappear.

But, why are statisticians so unhappy with us? When we collect data from biology, two major factors come into play – bias and biological variation. Such factors conspire to reduce both the reliability (precision) and validity (accuracy) of our data. This means that we often end up with noisy data capable of detecting mainly large

changes. Instead, statisticians want quiet, reproducible data capable of detecting small changes. This takes us to a largely unappreciated, but relevant point. Both bias and biological variation derive – at least in part – from reductionist theory and from the preferences of statisticians.

Although biology allows variation, it allows far less than what statisticians would lead us to believe. The isolated data favored by statisticians tend to maximize variation, whereas the connected data of biology does quite the opposite. Moreover, biology is entirely capable of supplying us with valid data, which, in turn, we can use to minimize the effects of the biases we create with our methods. In other words, if we want to, we can produce much quieter data.

Quiet data interest us here because they show the patterns, equations, rules, and algorithms biology uses to run its business. Since biology is in the business of optimizing outcomes in complex systems, access to quiet data gives us access to a wealth of proprietary information. As the story unfolds, we will learn to use such privileged information to our mutual advantage.

A few, brief examples will help to show where this story is going. First, however, we need to plant our feet on solid ground. Although most experts in academic and corporate circles identify biology as a descriptive science, complexity theory takes a decidedly different view. It prefers reality to convenience. Biology is a descriptive science now becomes biology is a quantitative science. By changing the definition, complexity theory compels all parts of our story to obey the mathematical rules of biology – even when we have no idea what they might be. Given this new reality, part of our mandate becomes one of finding and bursting the many bubbles created by the assumptions of a descriptive science. The example to follow shows how easy it is to get ourselves into deep trouble by bursting a bubble fundamental to experimental biology as it currently exists.

Most of us would agree that the primary goal of scientific studies is to detect changes and to explain why they occur. However, reporting changes in biology nourishes an enormous bubble. Why? Many laboratories and clinics collect data as concentrations, which, in turn, they use directly to look for biological changes. Recall that a concentration (A/B) includes two values, a numerator (A) and denominator (B). Drawing from our training in chemistry, we know that A can change, but B will remain constant because it represents a standard unit of volume that conveniently cancels out when the change is calculated. This gives us one value for the control (A_{t0}) and another for the experimental (A_{t1}) – everything appears to be in perfectly good order ($\Delta = A_{t1}/A_{t0}$). Here change (Δ) works.

When it comes to comparing concentrations, however, chemistry has one set of rules and biology another. In an experimental setting, we can expect chemistry to have two variables in play ($\Delta = A_{t1}/A_{t0}$), but biology with its added load of complexity will have four: ($\Delta = (A_{t1}/B_{t1}) / (A_{t0}/B_{t0})$) because $B_{t0} \neq B_{t1}$. In a biological setting, comparing concentrations produces uninterpretable results on a vast scale. Since most photometric measurements (i.e., optical densities) qualify as concentrations (Bolender, 2007, 2007A), even biochemistry contributes handsomely to the bubble when its data are related to a biological reference. This self-induced chaos is one of the enduring legacies of our descriptive science. Compelling evidence for the existence of too many variables in play appears throughout the literature as disagreements, inconsistencies, and irreproducible results.

In short, there are reasons for concern. As a complex and highly adaptive organism, we can adjust to even the harshest of research environments. Unfortunately, we may be reaching the limits of our endurance. A well-trained investigator with years of experience in the biological sciences is likely to produce a list of real-world hazards similar to the one given below. Our purpose here in preparing such a list is to assure the reader that all the items included

therein belong largely to the same problem. Moreover, the list serves as a convenient score card for the game we are about to play. The solution, as the book will explain, requires little more than sliding biology from one theory structure onto another – from reductionism to complexity. The list highlights the realities of our working conditions.

1. Acceptance of a descriptive science
2. Acceptance of a methods-driven science
3. Acceptance of faulty assumptions
4. Uncontrolled experimental bias
5. Uncontrolled biological variation
6. Uncontrolled false positives and negatives
7. Uncontrolled ambiguity
8. Inadequate theory structure
9. Inadequate research model
10. Inadequate publication model for research data
11. Inability to reproduce results routinely
12. Inability to detect biological changes reliably
13. Inability to quantify phenotypes exhaustively
14. Inability to deal effectively with biological complexity
15. Inability to correct methodological distortions
16. Inability to access biological information
17. Absence of first principles
18. Absence of data connectivity
19. Absence of objective diagnosis and prediction
20. Absence of mathematical markers
21. Absence of a universal database for published data
22. Absence of a common language shared with biology
23. Absence of published data compatible with biology

A word of caution is in order. This book is a hard read. The mere concept of biological complexity is still so far beyond our comprehension that most reasonable people avoid it altogether. To make matters worse, biology is only one part of a much larger problem. All of our methods for collecting and interpreting data contribute yet another level of complexity to that of biology. This means that gaining access to the core principles of living systems requires the unfolding of two interacting complexities - simultaneously (Figure P.2). Since this operation involves a monumentally tedious array of

details and arcane arguments, we will accede to treating complexity as a simple game that we can learn to play with biology – one move at a time.

Acknowledgments

The idea of approaching biology as a complexity came from a month long workshop held in Santa Fe, NM (1987) under the auspices of the Santa Fe Institute. It occurred 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 resulted in a strategic plan accompanied by a list of recommendations (Morowitz and Smith, 1987).

In turn, the insights and enthusiasm generated by this workshop led first 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 on going Enterprise Biology Software Project (2001-Present). The book summarizes the yearly reports of this project - all of which are currently available online (enterprisebiology.com).

In large part, the success of this project derives from the generosity of the stereology community in supplying reprints for the stereology literature database and to the Internet Brain Volume Database (Kennedy, et al., 2012) for providing online access to MRI data. Since many of the keys to understanding biology as a complexity already exist within the biology literature, we will use this book to show what our published data are capable of unlocking.

Introduction

What is a complexity game and why do we want to play it with biology? Biology plays the complexity game by translating its rules, procedures, and outcomes - stored largely in the genome - into phenotypes that can do extraordinary things. A phenotype represents a snapshot of an individual at a given point in time, linking the past to the present and the present to the future. By playing the complexity game with biology, we gain access to this phenotype along with a new strategy for interacting with biology.

The Problem

Biology operates as a complexity, wherein it defines and is defined by its parts and connections. In spite of this reality, we continue to study biology not as a complexity, but as a contrived simplicity. Our current theory structure operates on the assumption that we can take biology apart, understand the parts, and subsequently understand biology. The problem with this approach is that it lacks an appreciation for the order that comes from the connectivity of the parts and the emergent properties arising therefrom. Moreover, by exchanging reality for convenience we invite the penalty of unintended consequences.

Few realize, for example, that a theory structure based on reductionism limits our ability to create a mathematical foundation for biology analogous to those basic to physics and chemistry. By taking the complexity out of biology, we unwittingly abandon biology's connection to mathematics. This explains why biology remains a descriptive science. The underlying problem is one of dimensions. Reductionism, which eliminates complexity by removing its connections, also eliminates one dimension of the biological information. The remaining parts represent points (data) that now exist in zero-dimensional space. (Recall that statistical theory deals largely with the behavior of such data

points.) As a complexity, however, a living organism must operate in a dimensional space higher than zero because it must accommodate linear strings (patterns) consisting of parts and connections. The unavoidable truth is that biology, as an experimental science, operates on the risky assumption that we can use isolated information existing in zero-dimensional space to explain complex events occurring in higher dimensions. Abbot's delightful book (*Flatland*, 1991) offers a gentle introduction to the problem of information flow by describing what happens when we view the same world from different dimensions.

Since we can be reasonably confident that biology defines and executes its functions by rule, our main job here will consist of assembling a complexity parallel to the one of biology - using a more inclusive theory structure. We will discover that by restoring the complexity we can restore the mathematics along with many of its rules. This represents an import step because a quantitative approach allows us to play a far better game with biology.

The complexity game we are about to play must rank as one the most challenging. It comes without instructions and the user gets to determine the length of the game, the level of difficulty, and the size of the prize. When playing the complexity game with biology, however, it is up to the player - or players - to discover the rules and then figure out how to make the right moves on the right playing field. Experienced players have the distinct advantage of knowing that teaming up with biology all but guarantees a win. Biology already knows all of the rules, moves, and playing fields and seems perfectly willing to share this knowledge with us.

Theory Structure

The book introduces the reader to complexity by playing six games in order of increasing difficulty. Theory structure plays an important role

in that it guides the tasks of constructing the playing fields and figuring out to play a given game.

Figure 0.1 indicates that reductionist theory directs the first game, whereas the remaining games work together to assemble and test a new theory structure based on complexity. Notice in the figure that the first two games rely exclusively on the post-mortem data of biological stereology, whereas the remaining four use data collected with MRI from living subjects. This distinction is important because we will discover that a sharp line exists between these two data sources.

Two remarkable things will happen as we make the transition from the simplicity of reductionism to the complexity of biology. We will develop an unexpected confidence and skill in designing games of increasing complexity and, at the same time, take comfort from the discovery that the harder the game, the easier the solution. Lest we forget, however, our story begins at the point where we have absolutely no idea about how to study biology as a complexity or even if it is possible.

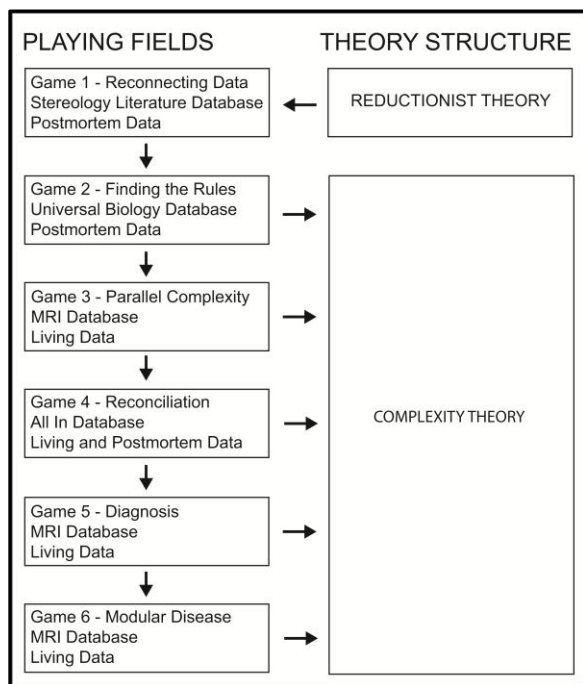


Figure 0.1 Playing the complexity game with biology. Notice that the games, which begin with stereological data derived from post-mortem samples, quickly progress to MRI data derived from living individuals. Each playing field consists of one or more relational databases.

Playing a complexity game requires meticulous attention to detail and a steely determination to recognize reality and play accordingly. Starting a complexity game with biology, however, can be a somewhat unnerving experience because all such games must begin in the altered reality created by reductionism.

Everyone knows that modern investigative biology plays largely by the rules of reductionism. This theory structure reduces the complexity of a living organism into a subset of isolated parts, but, at the same time, its methods quietly introduce artificial properties. Parts prepared for a stereological analysis, for example, may experience as many as thirty-three operations that can distort reality (Bolender, 2003) – along with the post-mortem consequences of going from living to lifeless. In spite of convincing evidence to the contrary, most biologists still consider reductionist data a valid representation of biology. This assumption, perhaps more than any other, deprives biology from enjoying the many advantages intrinsic to sciences based on first principles. Put simply, reductionist data will continue to make important contributions, but they will be largely ineffective in dealing with a host of pressing problems swirling around the real world of biological complexity.

Complexity theory takes its rules from biology as it normally exists and as it exists for us. Consequently, we will be dealing with two interacting complexities, one coming from biology and the other from the distortions we create by collecting data. Until the game advances to the point where these two complexities become separable, we will be playing with a handicap. The best we can do in the early games is to minimize the distortions and focus our attention on finding patterns in data collected with stereological methods. Such data are essential because they allow us to quantify biological parts

of all sizes and shapes in terms of volumes, surfaces, lengths and numbers. Moreover, morphological data represent the basic building blocks of a phenotype.

Games

A complexity game proceeds from one playing field to another by making moves that define the properties of the subsequent playing field. In its turn, each playing field supplies new forms of information with new data formats and patterns. Since patterns reflect underlying rules, they usually offer the best clues for figuring out what biology is doing.

In the text, a move begins with a question highlighted in blue and concludes with a color-coded answer - a green text box signals a win, red a loss. The intervening text includes the strategy behind the move and supplies the methods, results, and interpretations. If, as the game proceeds, you become lost or miss the point of an argument, you can always go back to the original papers, reports, guides, or software packages for help. (Note: Some of this information is available online at enterprisebiology.com) Many of the details related to the stereological methods of data collection and manipulation lie well beyond the scope of this book and can be found elsewhere (e.g., Weibel, 1979; Gundersen et al., 1988; Cruz-Orive and Weibel, 1990; Bolender et al., 1995; West, 2012).

Background

The central strategy of the project consists of extracting data from the biology literature and then using them to discover how biology operates mathematically. Since biological complexity resides in the volumes (V), surfaces (S), lengths (L), and numbers (N) of its parts and in their connections, stereology becomes the method of choice because it can estimate these parameters with unbiased sampling methods. In effect, stereology is ideally suited to the task of dealing with biological complexity – at all

levels – in both living and nonliving subjects. Moreover, it allows us to access the phenotype as a set of nested complexities existing in n-dimensional space.

Before the games can begin, however, we have to redefine our relationship to the biology literature. By entering stereological data into a relational database, they begin to lose their imposed isolation by becoming part of a large and coherent data set. The advantage of this new arrangement is that it allows us to look for local and global patterns in published data. Often, but not always, local will refer to the data of a single paper, group or individual, whereas global identifies data coming from many different papers, groups, or individuals.

Complexity consists of patterns that display mathematical properties. These patterns will first appear as absolute data (V, S, L, N) fitted to regression lines with coefficients of determination (R^2) equal to 0.9 or better (recall that as the R^2 approaches 1.0, data points distribute either on or close to their regression line). R^2 s close to one tell us that the relationship of one part to another suggests a mathematically defined order.

Whenever we collect data from biology, however, our methods invariably introduce uncertainty. Recall that stereological estimates carry an unknown burden of biases related to the preparation and analysis of biological samples – particularly when taken post-mortem. Although we can be confident that the design-based methods of stereology guarantee unbiased estimates derived from both living and nonliving sources, we can also guarantee that different sources – living and nonliving – can give different estimates for the same parts – depending on the distortions (biases) we introduce experimentally.

We will discover that one way of mitigating these experimentally induced artifacts is to form data ratios that can minimize the effects of the distortions. This strategy also makes sense mathematically because the data show that bi-

ology exerts a greater level of control on the ratio of its parts than on their absolute values. Adults of different sizes, for example, frequently display the same parts with different volumes, but similar ratios. From this, it follows that absolute values can be expected to exhibit more biological variation than when expressed as ratios. In short, forming ratios effectively minimizes distortions in our data produced by experimental methods and biological variation.

Notice that by replacing absolute values with ratios, we are following a deliberate strategy designed to take our cues directly from biology. We will discover that the rewards of such an approach can be considerable. By deferring to biology, it will do most of the hard work required to get us to our initial goal of constructing a parallel complexity – our proxy for biology as it actually exists.

For convenience, we will begin by defining the ratio of parts as a data pair (A:X:B) wherein two named parts (A, B) are connected numerically by the ratio (X:Y). (Note: dividing Y by X sets X = 1.) By reconfiguring the stereology literature database as data pairs, we obtain a universal biology database, wherein all the published data share exactly the same format. Operationally, this relationship of part (A, B) to connection (X:Y) defines a unit (i.e., an element) of biological complexity, one with universal connectivity. Given this more convenient data type, we will be able to find quantitative patterns practically everywhere we look.

Of course, searching for patterns in data aggregated from thousands of papers becomes a challenging and very time consuming task because the ratios (X:Y) supply continuous (i.e., analogue) values. This limitation will be easily overcome by assigning each data pair ratio to a decimal step (or bin) and then fitting these ratios to a regression equation ($Y=bX^a$), wherein the values of the exponent a and the coefficient of determination (R^2) both approach one. With such an arrangement, the power equations ($Y=bX^a$) approach linearity ($Y=bX$) and predict the original values with a maximum error not

greater than $\pm 15\%$. In effect, this reduces the stereology literature database to roughly 100 equations, wherein every data point defines and is defined by an equation.

These new decimal bins not only speed the task finding local and global patterns, but they also play a pivotal role in assembling the playing fields for our complexity games. Moreover, chaos theory provides some added cover. By shifting our data from an analog (continuous) to a digital (stepped) platform, we move them slightly away from their original order and toward the edge of chaos where they become infinitely more interesting and informative.

Notice the strategy in play. By translating the isolated data of individual papers into a large digital literature consisting of standardized data ratios and equations, our data can detect quantitative patterns and generate data sets large enough to qualify as a parallel complexity. By using these ratios to assemble playing fields of increasing complexity ($X:Y \rightarrow X:Y:Z \rightarrow X:Y:Z...N$), we can begin to attack difficult problems with surprising ease. Keeping everything on a mathematical footing keeps biology in the loop and allows us to benefit handsomely from our vast investments in basic and clinical research.

Several examples will serve to illustrate how data ratios provide a wealth of new information about the mathematical underpinnings of biology. Of special interest is the finding that biological parts and connections display valences and stoichiometries analogous to those found in chemistry. Biology uses the same strategy seen for elements and molecules by allowing the same two parts to form different ratios. This flexibility greatly increases the number of possible outcomes - including emergent properties. By increasing its potential for variation and adaptability, biology presumably improves its chances for success and survival. The same applies to us. By becoming privy to a strategy of such fundamental importance to biology, we find ourselves in a much stronger position to ask probing questions about how we currently collect and interpret our data. If biology want-

ed to give us a friendly nudge in the right direction, revealing its use of ratios and valences would be a clever way of doing it.

As the chapters unfold, we will discover how a given theory structure in biology comes with its own set of rules - often producing dramatically different results. Change in biology, for example, represents an enormously complex event, wherein a given part influences and is influenced by a large number of other parts and connections. In contrast, identifying a change in a single, isolated part ignores - almost entirely - the true nature of change in biology. Moreover, isolated data rarely contain enough information to get to the right answer. By looking at such truncated data through the lens of complexity theory, we can begin to understand why theory structure plays such an important role in the discovery process.

All games seem to involve an element of luck, and our complexity games are no exception. A chance encounter with an Internet database containing MRI data from human brains proved to be the game changer. It allows us to make key connections between theory structures (reductionism to complexity) and parallel complexities (living to nonliving). Moreover, by converting MRI data into mathematical markers, we can produce playing fields capable of diagnosing disorders of the brain objectively and begin to understand the role that quantitative relationships play in the disease process.

We will also discover that the brain uses many of the same parts and connections - acting as modules - to assemble a wide range of different disorders. Once again, we find biology reconfiguring itself to create new emergent properties - a theme repeating relentlessly at all levels of size. The big surprise is that these markers reveal a level of complexity so enormous that even the big data technologies of today may not be up to the task of explaining how these disorders appear and develop. Of one thing, however, we can be certain. The opportunities created by mathematical markers for triggering advances in our understanding of

biology are likely to surpass even our most optimistic predictions.

There is more. The MRI database of living brains can do something that the stereology database of post-mortem brains cannot. Only living brains are capable of displaying - routinely - identical patterns both locally and globally. We will use this remarkable property as an acid test for determining the validity of biological data.

Consider what this test will tell us. When we try to diagnose disorders in post-mortem brains using mathematical markers derived from their living counterparts, we will be disappointed consistently. This results from the fact that exactly the same parts in living and nonliving brains display different ratios and consequently different markers. We will use this inconsistency as an opportunity to identify and remove the distortions that exist in stereological data when they come from nonliving sources.

Challenges

Biology as a science faces a major challenge going forward in that it owns the responsibility of unraveling the complex relationships of genes to phenotypes. This means that stereology - with its extraordinary ability to quantify structure - becomes a critical player in working out the complexity of phenotypes because it can provide estimates for parts and connections of all sizes - seamlessly throughout the biological hierarchy. The immediate challenge for the stereology community becomes one of demonstrating - not just assuming - that equivalence exists between data sets derived from living and non-living sources.

Story

This book summarizes fifteen yearly reports (2001-2015) of the Enterprise Biology Software Project (Figure 0.2). Since these reports assume a working knowledge of biological stereology, readers unfamiliar with this method may miss some of the subtleties surrounding the forth-

coming games, moves, and interpretations. Consequently, care will be taken to explain what is going on behind the scenes.

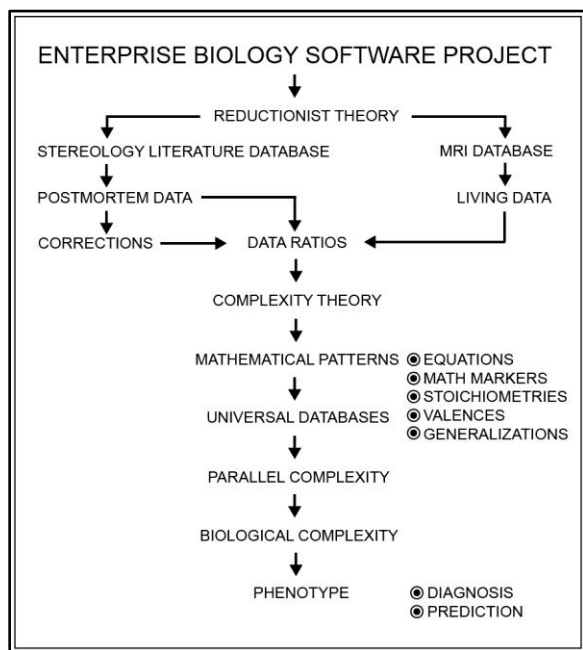


Figure 0.2 The Enterprise Biology Software Project explores the relationship of biology to theory structure to outcomes. It currently provides mathematical access to the phenotype and perhaps in time to the genome.

Before plunging into a seemingly endless parade of figures and arcane details, it may be helpful to begin with a brief summary of the complexity games and their outcomes. This offers the reader a quick overview of what to expect.

Perceptions

Our move into a theory structure based on biological complexity will introduce a number of fundamental changes in the way we think and operate. The following list offers a preview of these coming events.

1. Patterns replace individual data points as the primary source of biological information.
2. The biology literature – expressed as a universal biology database - becomes a single, global experiment to which each new publi-

cation contributes its data. In effect, the literature becomes the phenotype.

3. Data interpretation requires an active collaboration with large scale databases derived from the biology literature.
4. Experiments involve large numbers of connected patterns ($\sim 10^5 \rightarrow 10^7 \dots 10^n$) rather than small numbers of isolated data points ($\sim 10 \rightarrow 10^2$).
5. Change in biology is such a complex event that its interpretation requires collaboration on a massive scale.
6. Valences serve to define structural patterns in health and disease.
7. Data of the basic and clinical sciences interact seamlessly within in the framework of universal biology databases.
8. Decision-making derives from the collective knowledge contained within universal biology databases.
9. Playing fields define the games and their outcomes.
10. Gold standards for biological information come from living systems.
11. Parallel complexities operate on rules and algorithms consistent with the mathematical core of living systems.
12. Interpreting experimental outcomes typically involves the management of multiple complexities simultaneously.
13. Data from nonliving sources cannot be expected - a priori - to duplicate those of living ones.
14. The human brain in health and disease adheres to a modular design.
15. Theory structure influences experimental bias and biological variation.
16. The theories of reductionism and complexity combine to form a theory structure capable of supporting the biology enterprise.

Our current mindset in science revolves around the concept of variables (x, y), which relate to one to the other by some function ($f(x) = y$). In contrast, biology seems to prefer patterns expressed as numerical ratios, which it uses to generate complexities and emergent properties. Consequently, we will explore the ratio-

based patterns of biology with stoichiometries, valences, Fibonacci series, harmonies, design codes, polynomials, mathematical markers, modules, graphs, and cluster analyses.

Bigger Picture

We - as a scientific community - live in a world constructed as a simplicity, wherein our biological information consists largely of disconnected elements. Biology, on the other hand, lives in a complex world wherein these same elements exist in a highly connected state. Although we often recognize this inconsistency, we seem perfectly willing to accept the ways things are. This, of course, imposes limits on what we can do.

By playing the complexity game, we address two compelling questions. How do we go from state A (simplicity) to state B (complexity) with a minimum amount of discomfort and is making such a trip really worth the effort? In attempting to answer these questions, we will be putting ourselves in a curious position. We will have to decide if we are going from fantasy (A) to reality (B) or just from one fantasy to another. Making such a distinction will require a new type of information, one produced by combining the data and expertise of thousands of our best scientists into constructs capable of addressing real world problems. In effect, we will have to follow the data to wherever they lead.

Fortunately, the games, which are driven by data and often risky moves, will flush out more than a few startling surprises along the way. Biology - as a neutral referee with impeccable credentials - will serve as the presiding judge. Such an arrangement requires that we take a first principles approach, one that puts everything to the test by biology and by anyone else willing to try.

The deep understanding to come from this exercise results in a paradox. Complexity turns out to be far simpler than simplicity - because complexity runs on a mathematical platform,

whereas simplicity does not. If we are prepared to listen, this is what biology is about to tell us.

Language

For a science to work properly, it must include a two-way communication system capable of exchanging information between the parties involved. In physics and chemistry, such a system exists. We use first principles embedded in a theory structure to interact mathematically with the physical world. Although the science of biology uses a similar theory structure (reductionism), it largely lacks first principles, does not speak mathematics, and ignores complexity. By failing to interact mathematically with the natural world, we create a language barrier that prevents us from advancing beyond the level of a descriptive (soft) science.

If, as suggested by Adami (2015), we can define life as information stored in a symbolic language, then our story about biological complexity also becomes a story about language. By following the data, we will find that phenotypes can in fact be translated into a symbolic language consisting of pieces of information (one-dimensional strings) that connect in n -dimensional space, where $n \geq 1$. Using databases populated with strings numbering in the millions, we will discover - much to our delight - that we can communicate with biology objectively. When we pose a question quantitatively, biology politely responds with an answer.

The strings of our symbolic language, which include alpha names and numeric ratios, identify patterns that relate our descriptive names for biological parts to the quantitative properties given to them by biology. Since biology uses quantitative rules to form patterns, we will do the same. By allowing our strings to grow from characters (mathematical markers) to words to sentences to stories, we will be duplicating the syntax of biology. In turn, these linguistic strings give us a theory structure based on complexity, detect first principles, speak mathematics, and allow us to set up and solve deliciously difficult problems.

Chapter 1

Game 1 – Reconnecting Data

The first game takes its inspiration from a month long think tank held under the auspices of the Santa Fe Institute, which is summarized in a report by Morowitz and Smith (1987). The charge given to the more than fifty participants was to figure out how to organize all of the data of biology, thereby encouraging new connections, theory structures, and discoveries.

Complexity, as we all know, consists of many parts and connections with local, global, and emergent properties – all subject to rules. If we approach biological data as a complexity, then it follows that all the many parts and connections must be quantitative and hierarchical. The only biological method capable of delivering such a wide-ranging data set is stereology. Recall that this approach uses design-based sampling to estimate the volumes, surfaces, lengths, and numbers of biological parts of all sizes. Herein we find the argument for designing, populating, and testing a biology literature database populated with stereological data.

The first game sets out to reconnect the isolated data of the stereology literature, first by storing them in the same place and then by allowing them to interact. Note that each game begins with a goal followed by several moves intended to achieve it.

Move 1: Can we organize biological data within the framework of a relational database?

The purpose of the move was to take data from highly heterogeneous sources (research publications) and standardize them. A primary requirement was a database model capable of accommodating a majority of biological data – including both structure and function.

This required access to thousands of reprints, many of which were generously supplied by members the stereology community. The Enterprise Biology Software Project was set up as a vehicle for developing new technologies and returning them along with a yearly report to contributing authors. Currently, the project supports investigators working in more than forty-five countries.

1-1 Relational Database

The relational database model includes a structural hierarchy consisting of sixteen compartments (hard-coded), twelve structural data types (hard-coded), and three functional data types (user-defined). It also includes tables for authors, citations, and methods (Figure 1.1). Notice that the database model uses two structural hierarchies – one for control data (co) and the other for experimental (ex) - connected to one another and to either a control or experimental data table.

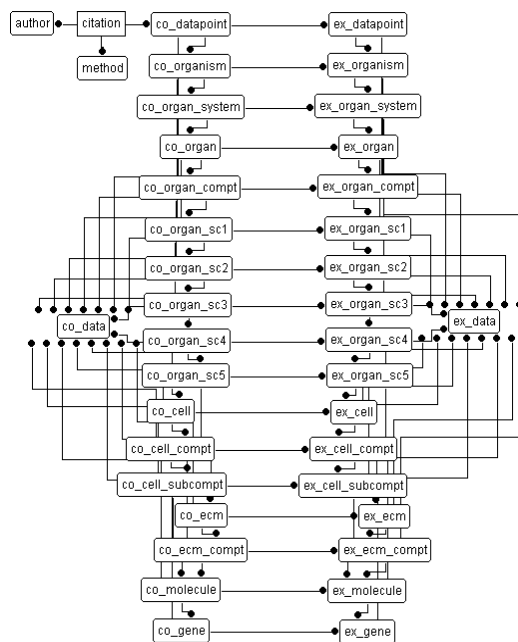


Figure 1.1 The logical database model for the biology literature includes a collection of entities (boxes) and relationships (lines), as defined by rules of relational databases (From Bolender, 2001A).

In turn, the logical model of Figure 1.1 becomes a physical model (Figure 1.2), wherein entities include the columns of database tables and relationships the joins between the tables. The user interacts with the database through data entry forms, browsers, simulators, and query screens.

Figure 1.2 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).

1-2 Entering Data

Data entry consists of first building a structural hierarchy for each data point and then mapping numerical data to it. Data expressed as volumes, surfaces, lengths, and numbers can be related to a unit of volume (concentration or density), to a structure, or to an average structure. Data entry includes extracting data from publications (Figure 1.3), standardizing data (Figure 1.4), and harmonizing units (Figure 1.5).

Figure 1.3 A surprisingly large number of publications report data exclusively as graphics. This work screen simplifies the task of translating graphical data back into numerical values (From Bolender, 2001A).

Figure 1.4 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 a majority of authors. The green screen serves as the template for data entry (From Bolender, 2001A). Terms and definitions appear at the right.

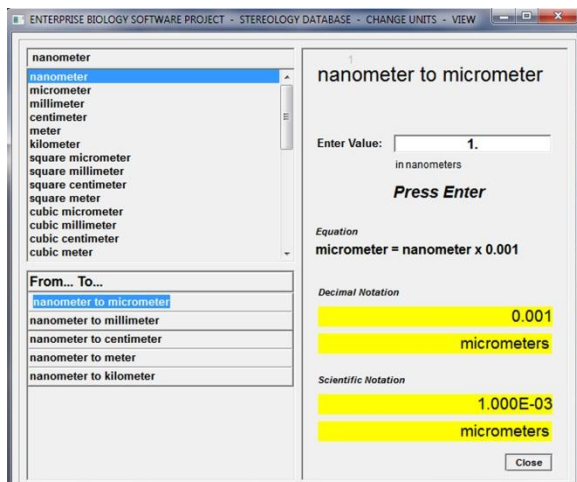


Figure 1.5 The units screen simplifies the task of converting from one unit to another (From Bolender, 2001A).

1-3 Data Catalogues

The logical database model of figure 1.1 uses the hierarchical relationships of biological data to map location to exposure. By organizing the literature into a single system of connected data, the model allows us to transform the data set into new formats or catalogues – as the need arises. For example, Figure 1.6 illustrates that we can view the literature one paper at a time (top), hierarchically as individual tables (middle), or as total data tables (bottom). More importantly, relational databases completely change our relationship to published data. Instead of remaining static, data become dynamic and capable of creating new forms of information.

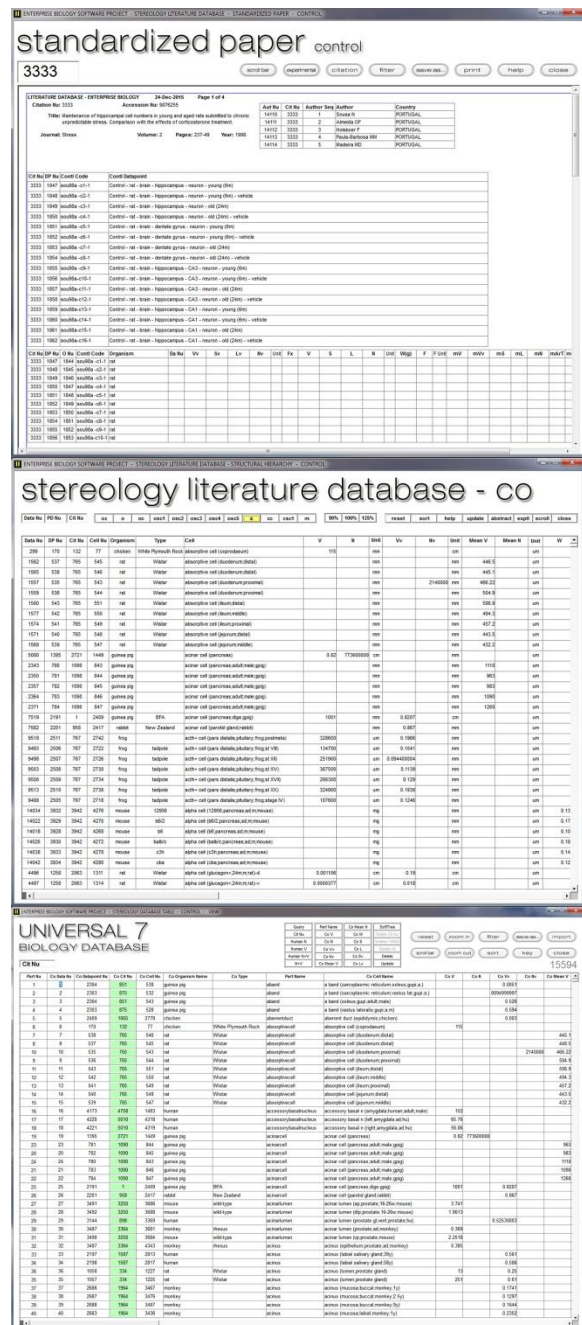


Figure 1.6 When stored in a relational database, the biology literature becomes a catalogue of data that can be expressed in a variety of configurations (From Bolender, 2001A).

Move 1: Can we organize biological data within the framework of a relational database?

Yes, we can extract biological data from refereed publications, organize them, and store the information hierarchically in a relational database. In a digital format, biological data increase their value importantly by generating new forms of information.

1-4 Reductionist Theory

Collecting and interpreting data fall within the purview of a theory structure, which for our purposes in this chapter follows the rules of reductionism. Reductionist theory holds that a complex system equals the sum of its parts. When applied to biology, it reduces complexity to simple terms by isolating and studying the properties of individual parts.

1-5 Playing Fields

The stereology literature database serves as our first playing field. It represents - in digital form - data coming from thousands of refereed publications selected with regard to the methods and to the perceived ability of the data to detect biological changes (Bolender, 2001a, 2007). Since we now have a playing field, we can make our first move with the database. Recall that a move, which takes the form of a question, occurs on the playing field according to the rules of the theory structure in play. In a reductionist setting, we know that both concentrations and absolute values are used routinely to detect biological changes, but we know little about the consequences of obeying this rule. Move 2 puts the question.

Move 2: Do both concentrations and absolute values detect biological changes similarly?

Reductionist theory assures us that we can detect a change in a given part independent of biological complexity. Were this true, we would expect to find wide spread agreement of experimental results across the biology literature. In fact, quite the opposite occurs. The literature overflows with conflicting results and all too often studies cannot be repeated. Since a biological change represents a complex event, many variables will be in play that can influence the outcome, which includes the unintended consequences of our experimental methods. Move 2 identifies a bubble (concentration) and estimates its cost to the community.

1-6 Concentrations

Try this. Pick up a recent weekly journal and thumb through the biology articles. You may discover that most of those studying biological parts will report changes by comparing concentrations, often expressed as some version of an optical density. Under reductionism, this method of detecting biological changes is perfectly acceptable. We can check on the appropriateness of this practice by viewing the same changes with and without complexity.

If both concentrations and absolute values allow us to detect changes, do they detect the same changes? If yes, then the results of one should equal the results of the other. Stated mathematically, the question becomes:

$$\frac{\text{Concentration Value}(E)}{\text{Concentration Value}(C)} = \frac{\text{Absolute Value}(E)}{\text{Absolute Value}(C)} \quad (1.1)$$

where C equals control and E experimental.

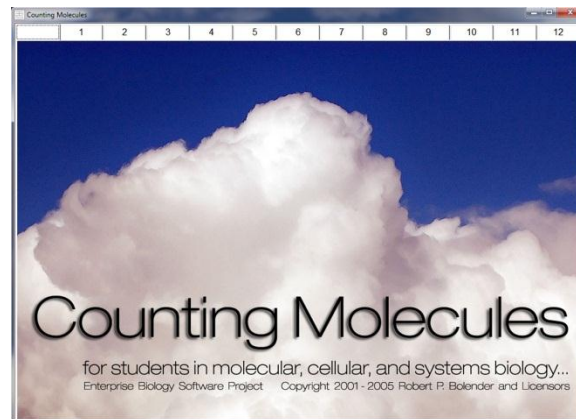
Since the stereology literature database offers ready access to both concentrations and absolute values from the same papers, the playing field (stereology database) answers the ques-

[illegible]

On average, the two estimates – concentrations and absolute values – agree only about 50% of the time. This discrepancy begins to explain the risk involved in simplifying biology by throwing away its complexity. The problem, however, goes much deeper. Notice that the concentration trap identifies an ambiguity, but only begins to resolve it. Since estimating absolute values makes more sense mathematically, we now use them widely to detect biological changes – assuming that any remaining ambiguity is unimportant. Of course, such an assumption can be dangerous, as we will discover later in our story. More importantly, however, the persistence of ambiguity in our research data signals an inability of our methods to manage risk effectively.

- typically a cubic unit of reference volume - contains the parts of interest from exactly the same number of cells. When not the case, a change the number of cells contained within the cubic unit of reference can influence the result or produce a change entirely by itself. Since cells routinely change their shapes and volumes in response to experimental exposures and the methods of preparation, comparing concentrations represents a high-risk approach to detecting biological change. From a mathematical standpoint, comparing concentrations in a biological setting involves the behavior of four variables not the widely assumed two.

The Counting Molecules program (Bolender, 2005) simulates a wide range of experiments routinely encountered in biochemistry and stereochemistry (Figures 1.7 and 1.8). By expressing a biological change in terms of a set of interconnected variables, one can quickly discover the importance of connectivity when looking for changes in biological parts. Figure 1.8 shows that the same experiment can give different results when run with and without complexity.



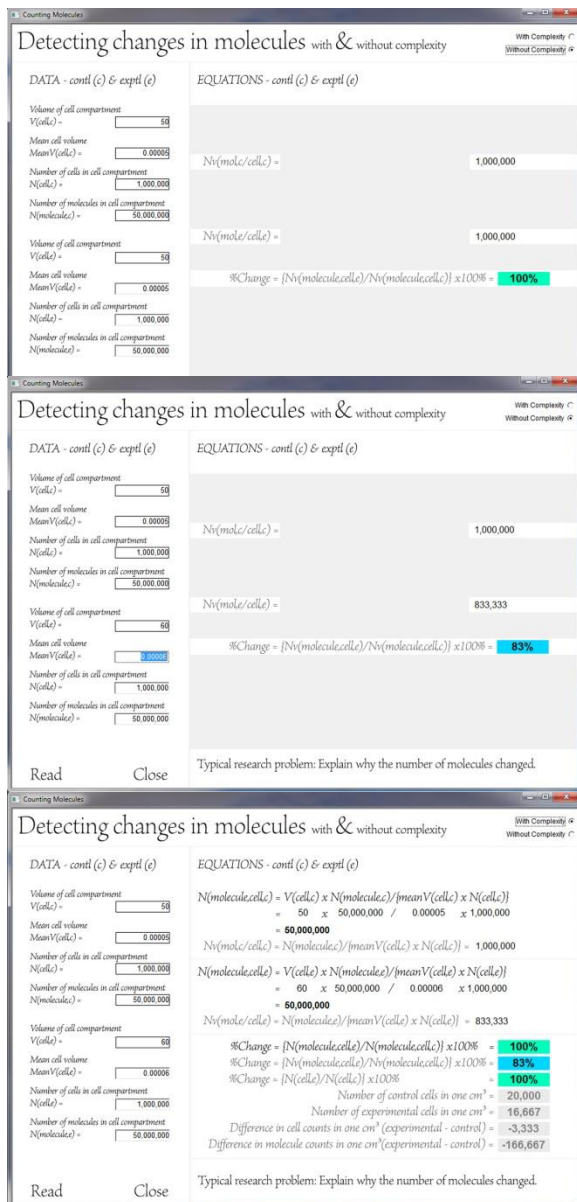


Figure 1.8 Detecting changes in molecules by comparing concentrations (i.e., optical densities) carries a risk. The top panel shows that the concentration of molecules – expressed as a numerical density (Nv) – is the same (100%) for both groups at the beginning of the experiment (1,000,000/cm³). After running the experiment, the middle panel 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. Most studies end here. The bottom panel 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. In fact,

the absolute number of molecules remained unchanged. Isolated variables (top and middle panel) deprive the investigator of the critical information needed to interpret the data correctly (From Bolender, 2005).

1-7 Absolute Values

When reporting experimental outcomes, switching from concentrations to absolute values improves the outcome by removing a major source of ambiguity. To do this, we multiply a concentration (N/V) by an absolute volume (V). The stereological equation for this operation is given as:

$$N_{\text{molecules}} = V_{\text{part}} \times N_{\text{molecules}} / V_{\text{part}}, \quad (1.2)$$

where $V_{\text{part}} = V_{\text{part}}$.

Note that equation (1.2) requires that the contents of a cm³ of the absolute volume (V_{part}) are identical to the cm³ of the concentration (V_{part}). If not, then the two cm³ will not cancel out – apples and oranges. Whenever we assume that unequal things are equal, we create a bubble.

The hierarchy equations of stereology designed to solve for absolute values often involve multiple stage sampling. This increases the likelihood that the cancelling rule cannot be obeyed. When the sampling stages combine data from fresh and fixed tissues or from two types of microscopy (light and electron), the chance of encountering identical cm³s across the multiple stages may be close to zero. Consequently, implicit in the absolute values estimated with hierarchy equations is the assumption that all the cm³s cancel – even when it may be both theoretically and practically impossible. Solutions to such problems of data management cannot be accommodated within the framework of reductionism because the theory structure is not designed to deal with complexity. Recall that the whole point of reductionism is to eliminate complexity.

Consider a typical hierarchy equation with data coming from fresh tissue, light microscopy, and electron microscopy:

$$N = \text{fresh tissue} \times \text{light microscopy} \times \text{electron microscopy}$$

$$N = V_{part2} \times V_{part1}/V_{part2} \times N/V_{part1} \quad (3)$$

We must accept two highly suspicious assumptions:

$$V_{part2} = V_{part2} \text{ and } V_{part1} = V_{part1}. \quad (4)$$

In the absence of corrections for our methods-induced distortions, one can argue that the experiment is primarily interested in detecting a change not in getting the most accurate estimates. When, however, we commit to this argument, a new set of assumptions come into play. Now the distortion in the volumes seen in the controls must be identical to those of the experimentals. Otherwise, the volume distortions will disrupt the two estimates unequally and produce an unstable outcome. Once again, we are forced into making another risky assumption.

Regrettably, experimental biology suffers from a fundamental problem. Reductionist theory allows us to detect changes in isolated parts by comparing either concentrations or absolute values. To do so, however, we must make assumptions that often appear inconsistent with reality. Nonetheless, reductionism as a theory structure remains largely unscathed because of the isolation it creates. By eliminating complexity, it also eliminates our ability to hold it accountable. Clearly, such an invincible design represents a brilliant construct (attributed to René Descartes, 1596-1650).

However, such brilliance can have wide reaching consequences. Recall that one of the most respected ways of demonstrating correctness in experimental biology derives from the notion of reproducibility. If several different laboratories do the same experiment and get the same result, then the outcome would seem to be correct. Implicit in such a conclusion is the assumption that the data are correct to begin with. If this is not the case and the data are incorrect, then a finding of reproducibility obviously leads to the wrong conclusion. In fact, reproducibility confirms precision not accuracy.

Game 2 showed that concentrations and absolute data can detect changes in the same parts differently (Figure 1.7), but it told us nothing about how these estimates were affected by biases and experimental errors. It did warn us, however, to expect consequences when we remove complexity from our data.

Move 2: Do both concentrations and absolute values detect biological changes similarly?

The answer to move 2 is no. On average, concentrations and absolute values give the same results only 50% of the time.

Since detecting changes reliably is a mission critical requirement of any science and since reductionism appears to fall short of that goal when applied to biology, it may be time to rethink our approach to change. Accordingly, our next move pursues a new strategy.

Move 3: Can we identify routinely quantitative patterns in biological data?

To make this move, we turn once again to the stereology literature database for help. Patterns, you may recall, lead to generalizations and generalizations to rules (Figure 0.2). In turn, rules combine to form new theory structures.

Biological data tend to be noisy for a variety of reasons. We marginalize our data by adding biases, accepting high levels of biological variation, trying to detect biological changes with concentrations, and failing to enforce unbiased sampling rules. Stereology offers a significant advantage because it is designed specifically to eliminate several of these troublemakers.

1-8 Design Code Equations

Perhaps the easiest way of finding patterns is to fit biological data to curves with regression analysis. The best patterns tend to include regressions with coefficients of determination (R^2) close to 1.0. Such an orderly array of variables suggests the presence of underlying rules, presumably exercised by biology. Since the stereological literature database contains a rich source of readily available data, hunting for such patterns becomes a relatively simple task. We can plot everything against everything else – controls vs. controls, controls vs. experimentals, and experimentals vs. experimentals, using data from one or several papers.

This plotting exercise produces a large collection of regressions (design code equations), which when stored in database tables simplify the task of looking for generalizations in one (Figure 1.9) or many papers (Figure 1.10).

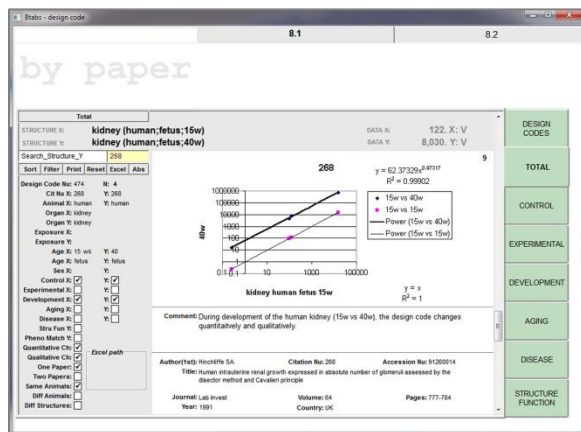


Figure 1.9 Local changes during development of the human kidney identify a distinct ordering of parts (Adapted from Hincliffe et al., 1991; From Bolender, 2003).

Figure 1.9, for example, shows that parts of the kidney grow proportionately, as suggested by the almost parallel relationship of the standard curve ($Y=X$) to the experimental ($Y=62.37X^{0.97}$) – the slopes of both curves are similar ($1 \sim 0.97$).

Figure 1.10 plots control versus experimental data for parts of the lung taken from three animal species – guinea pig, pig, and rat. Notice that the three animals share the same curves

and apparently follow the same set of rules related to change. Be aware, however, that these curves display R^2 's close to 1.0 (0.999) because only those points sitting on the curve or very close to it were included in the analysis.

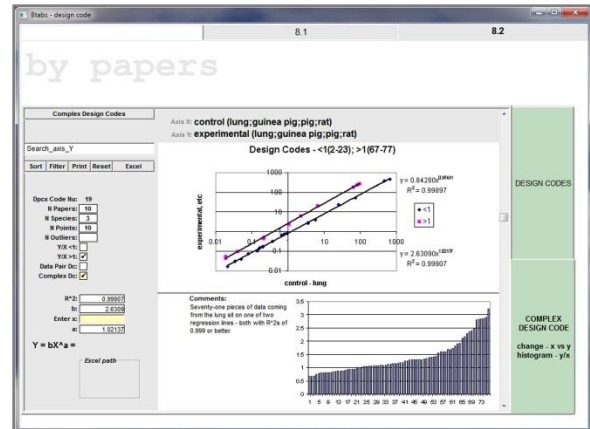


Figure 1.10 Global changes in the lungs of three different animal species taken from 10 papers (From Bolender, 2003).

When we plot a control part against its experimental counterpart, it may or may not show a change. In the absence of change, the control (X) and experimental (Y) values will be the same, and the equation will be linear: $Y=X$. In Figure 1.10, the curve passing through the origin with $X \sim Y$ and a slope ~ 1 ($Y=0.84X^{0.98}$) becomes the candidate for no (or little) change and the one displaced upward for change ($Y=2.6X^{1.02}$). Notice that the two curves tend to parallel one another because they have similar slopes (0.98 vs. 1.02). The figure shows that a change can change the amounts of a given set of parts, but not their proportions and that the same change can occur in different species. It also hints that the rules for producing a change can generalize across several animal species – at least for a specific set of parts in the lung.

By looking at many plots of biological parts undergoing changes in a variety of settings (Bolender, 2003), the patterns lead to generalizations, as summarized in Figure 1.11.

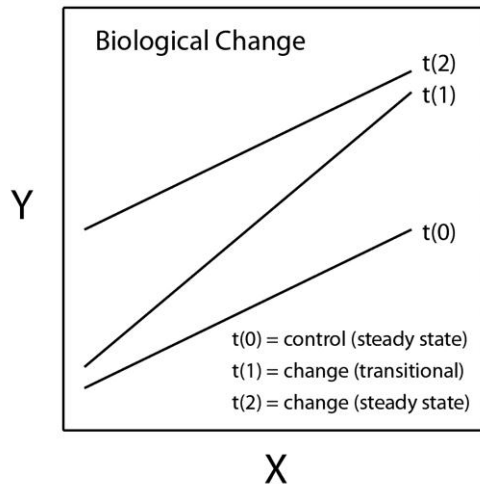


Figure 1.11 Change in biology can occur in well-defined steps, where Y = experimental parts, X = control parts, and t(0) → t(1) → t(2) (From Bolender, 2003).

A biological change, when expressed by a given set of parts with $R^2 \sim 1$, often cycles through a series of distinct steps as the process continues (Figure 1.11). As the change proceeds from no change (time 0) to change (time 1), the parts increase in amount at different rates and the curve becomes nonparallel (time 2). In time, the original proportion of parts becomes reestablished and the new curve (time 3) becomes parallel to the original one (time 0). In effect, change occurs as a highly choreographed series of events, wherein the phenotype undergoes a modification in response to a complex cascade of genetic events. Although the same ratios persist at t(0) and t(2), they become fluid at t(1) when the growth algorithms are in play.

The curves shown in Figure 1.11 tell us that we can expect different patterns of genetic expression at t_0 , t_1 , and t_2 . This may become useful to know when we want to map – in detail – phenotypic changes to genetic events. By using the phenotype to determine when and where to look and at what, we can improve our chances of a successful outcome.

Move 3: Can we identify routinely quantitative patterns in biological data?

Yes, design code equations suggest that quantitative order – detected as equations with $R^2 \sim 1.0$ – exists throughout the stereology database. They tell us that biological parts exist not in isolation, but rather in well-defined relationships – one to another. Change can alter these relationships, but often only temporarily.

1-9 Summary of Chapter 1

Our first three moves have taught us several things. We can deliver a relational database model for the biology literature, one that changes our relationship to published data. Instead of being isolated on the printed page, data are now free to form new relationships and patterns – with surprising ease.

By storing our data in databases, we begin the process of challenging the assumptions of reductionist theory. By taking biology apart and looking for changes in individual parts, we accept or ignore – often wittingly – a host of risky assumptions. Our methods, which allow us to detect changes in a variety of ways, tell us little about what produces a change. Instead, we assume – more often than not – that all the changes we observe can be attributed to biology and that the influence of bias and experimental error is of little importance.

Game 2 will attempt to deal with some of the obvious shortcomings of our current theory structure – reductionism – and begin to mitigate the effects of bias and biological variation.

Chapter 2

Game 2 – Finding the Rules

Reductionist theory is a construct of the scientific community, one used universally in physics, chemistry, and biology. When, however, we apply a rigorous mathematical method (stereology) to biology in a data-driven environment, uncertainties begin to surround the methods and results produced by reductionist theory (Game 1). Since these limitations are both real and self-evident, we need to consider a theory structure more responsive to biology with its pervasive and interacting complexities.

Move 4: Can we identify the properties of a new theory structure for biology?

2-1 A New Role Model

Move 4 poses a challenge because to come up with a new theory structure for biology we have to anticipate everything at a time when we know practically nothing. The only confidence building approach to such a riddle is to get help from a reputable source, one that knows practically everything and must anticipate little.

If we recruit biology and let it construct the new theory structure for us, then we are back in business with a first rate player in charge. This releases us from the otherwise inescapable burden of presuming to know something impossible to know at the outset. By letting biology do all the heavy lifting, the difficulty of our task reduces to keeping an open mind, watching, and gathering the many little details needed to play the complexity game. As we proceed, this strategy of going to biology for help will repeatedly reward us with winning outcomes.

We begin with what we already know. Recall that the design code equations were able to

deliver regressions with R^2 's close to one because the relationships of one part to another – all along the line - represented ratios. We also know – by definition - that a collection of related parts and connections defines a complexity.

If we assemble a playing field using the same parts and connections that combine to form biological complexity, then our new theory structure anchors itself securely into the mathematical bedrock of biology.

What we do not know at this point, is that a theory structure taken from biology will give us a complexity parallel to the real one. How will this be helpful? Biology plays very subtle, but extremely high-level games of which we know practically nothing. It, for example, routinely uses complexity to trigger a vast array of emergent properties. These properties interact with their surroundings by instigating changes and feedback loops. By combining a careful attention to detail with a generous assist from biology, we may eventually learn to do the same with our own parallel complexity.

Game 2 begins by assembling a playing field similar to the one biology uses to define its structure. The basic building blocks, which include parts and connections, combine to form complexities both locally and globally. Once again, we will recognize such complexities as patterns, which we can capture numerically with ratios, equations, and graphics.

2-2 Data Pairs (Ratios)

If organization is the first step toward understanding complexity, then storing data from the biology literature in a relational database starts the process. The hard lesson learned from move 3 (Chapter 1) was that access to a large

collection of concentrations and absolute values did not translate into an over abundance of quantitative patterns. At the time, it appeared that the putative patterns were most likely being overwhelmed by methodological bias and animal variation and underwhelmed by the relatively small sample sizes. In other words, the data were probably too noisy and too few in number to capture biology as it actually exists. In other words, the rules were beyond our reach.

Accordingly, the focus of the project shifted to finding ways to minimize the effects of methodological bias and biological variation and to increase the sample size (Bolender, 2001-2004). Forming ratios of one part to another improved the data in several ways. When two parts form a ratio, the effects of biases common to both values effectively cancel out. However, when this is not the case, the unshared biases remain and continue to distort the value of the ratio.

Although we can identify many potential sources of bias (Bolender, 2002; 2007A), making meaningful corrections requires an ability to separate biology from the artifacts created by our methods – a skill not yet available to us. Nonetheless, forming ratios reduces the effects of animal variation coming from different sized animals with different sized parts - within and across species. Recall that biology allows absolute values for its parts to vary widely, but is much stricter in maintaining the ratio of one part to another. This preference of biology for ratios identifies a first principle.

Shifting from data points to data pairs eased the sample size problem - substantially. Starting with the data set of a given paper, data pairs were formed by generating all possible permutations – taking two parts at a time. This quickly produced a data pair table containing more than 50,000 entries (Figure 2.1). This new database - based on ratios – became the playing field for our first complexity game.

2-3 Universal Biology Database

As the name implies, a universal database can store all types of biological data as ratios in a database table. Data used to form these ratios can include volumes, surfaces, lengths, and numbers, along with most data derived therefrom (Bolender, 2005). Notice what happens. By storing data from many sources in the same place, they can work together across disciplines, animals, and settings. In effect, the database maximizes the likelihood of finding patterns in published data.

Initially, this “data togetherness” approach may seem somewhat curious in that our training teaches us quite the opposite. When we collect data from biology according to reductionist theory, we dutifully isolate them from their normal surroundings and then set them apart from their peers by our method of publishing. When, instead, we store data as data pairs in a common database table, they can connect, form patterns, and begin to tell stories.

Move 4: Can we identify properties of a new theory structure for biology?

Yes, the new theory structure will treat biology as a complexity, consisting of parts and connections. A ratio of parts will serve as the basic unit of complexity and populate the tables of a universal biology database.

Finding patterns, however, can become problematic - particularly when data are sparse and disconnected. One way of alleviating this problem is to assign each data point to an equation. Move 5 uses the data pairs in the universal biology database to explore this approach.

Move 5: Can an equation predict data points and a data point an equation?

Consider the table in Figure 2.1. It includes a collection of data pairs expressed as ratios $X:Y$. If we divide Y by X , X becomes 1 and Y becomes some number i , $X:Y = 1:i$. If we sort the table on column i in descending numerical order, we can continue fitting the Y values to a regression line ($y=xb^a$) until the coefficient of determination (R^2) reaches and maintains a value of 0.9999. When the R^2 starts to fall below 0.9999, we stop, back up slightly, and start a new regression equation for the next step. This process continues until all the data points (50,000+ rows in the database table) belong to a regression equation. Now each data point predicts an equation and an equation data points. By dividing the predicted value for Y by its observed value (the original number), we can see that the equations predict the outcomes remarkably well (Figures 2.1).

[illegible]

Figure 2.1 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).

Move 5: Can an equation predict data points and a data point an equation?

Yes, by forming the ratios $X:Y$, wherein X is set equal to one (Y/X), a list of ascending values can be fitted to regressions with an $R^2 \sim 1$. By relating each data point to an equation, we have a workable solution to the problem of sparse and disconnected data.

These new expressions, called repertoire equations, generate a wide range of patterns, solutions, and insights. Move 6 explores the potential of these equations with their data pair ratios. Notice that we have begun the process of gathering evidence to convince ourselves that biology is operating - by rule – and running on a mathematical platform.

Move 6: Can we extract patterns as equations from a table of data pair ratios?

A universal biology database made up of data pair ratios begins to change our perception of the biology literature in several ways. Published data go from inaccessible to accessible, disconnected to connected, passive to active, and inflexible to flexible. These improvements come from merely watching biology and duplicating the way it orders its parts - quantitatively.

Our universal database also allows us to discover that biology's rules will define much of what we can do. Data pair ratios, for example, can detect these rules as equations that begin to explain how biology uses its parts and connections to create complexities. The problem, however, is that our biological data carry many types and levels of complexity that we will have to figure out how to identify, access, and interpret.

2-4 Repertoire Equations

Making the transition from reductionism to complexity requires little more than switching from concentrations and absolute values to dimensionless ratios. Repertoire equations, which plot data pairs (X vs. Y) as regressions, detect widespread order as patterns within and across publications (e.g., Figures 2.2, 2.3, 2.4). They show how biology – as a complexity – manages its parts.

Figure 2.2 indicates that different cytoplasmic organelles form well-defined ratios with the endoplasmic reticulum and that the same ratios can persist within and across species.

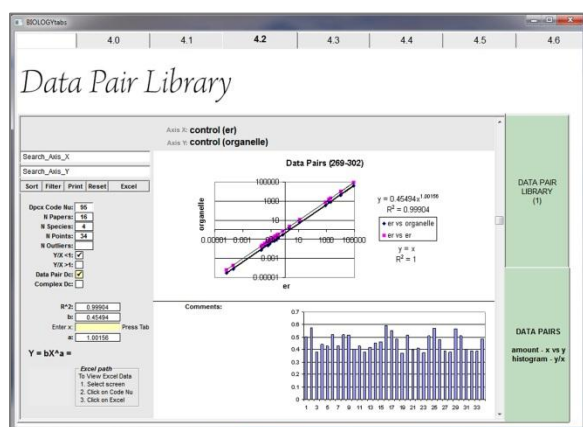


Figure 2.2 A high degree of order exists in the relationship of the endoplasmic reticulum to other cell organelles (From Bolender, 2004). The figure demonstrates that different organelles from different sources can occur in exactly the same proportion with the endoplasmic reticulum. Notice that the experimental curve (dark line) is linear (the slope=1.0), which explains why it can be parallel to the reference line (er vs. er) (From Bolender, 2004).

Extracting patterns from relatively sparse data sets, however, remains a challenge. Typically, the immediate result of plotting two different parts against each other are clumps of points with R^2 far from removed from 1.0 (Figure 2.3; Bottom: before). However, the clumps typically contain an underlying order (same figure, after - top) that we can extract with repertoire equations. These equations begin to explain the way biology constructs itself using rule-based ratios.

The equations also tell us that a given part can share connections (ratios) with many other parts and that the same two parts of a pair can connect to one another in different proportions. In other words, parts larger than molecules display valences analogous to those of chemistry. A potentially confounding property of valences is that they can vary widely when a change is occurring (Figure 1.11 – (t1)).

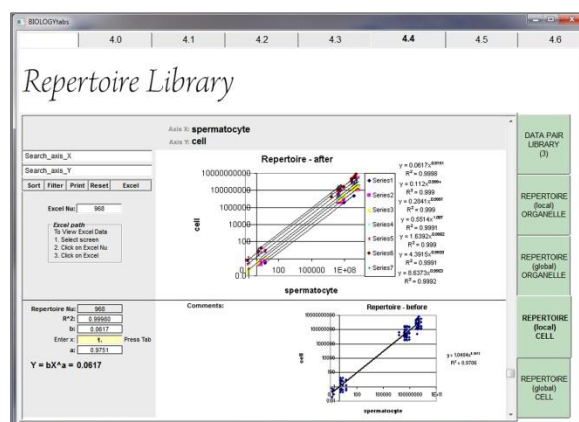


Figure 2.3 Regression analysis turns clumps of data (Repertoire - before) into sets of parallel repertoire equations (Repertoire - after) (From Bolender, 2004). Notice that the many different cells of the testis show a remarkable degree of order.

This newfound ability to transform noisy data (clumped) into quiet (equations with $R^2 \sim 1$) occurs often enough to create a repertoire library of examples (Figure 2.3). Moreover, the process works throughout the biological hierarchy of size.

A worked example explains the method. We begin with estimates for the volumes of Golgi (X) and mitochondria (Y) collected from a variety of animals (Bolender, 2004). Go to the data pairs table (Figure 2.1), type <mito> into the x name field, press Enter, click on sort Y button, and save the results to an Excel file. Now in Excel, start with Golgi and sort the X/Y column (containing the numerical values) low to high. Next, highlight the first three data pairs of Golgi and select a scatter graph. Change the axes to logs and fit the three points with a power regression line. If the R^2 does not approach 0.999,

add extra points – line by line – until it does. When the R^2 comes close to 0.999, stop and start a new regression. The result is a set of parallel repertoire equations defining the quantitative relationship of mitochondria to Golgi (Figure 2.4).

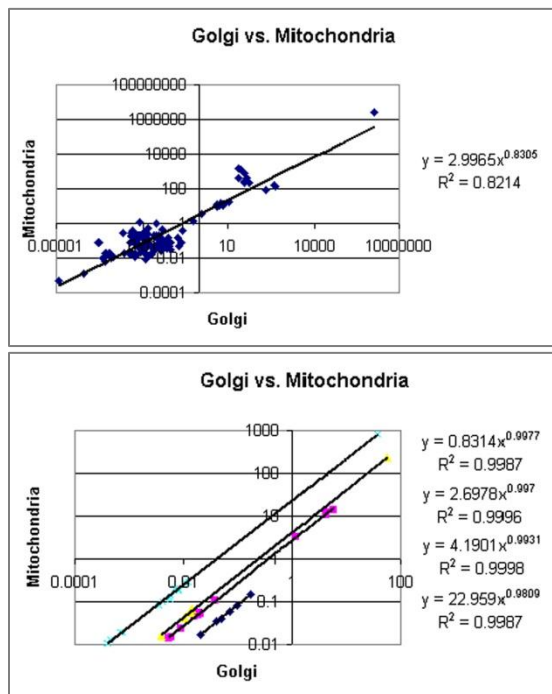


Figure 2.4 Top: A repertoire plot shows the relationship of mitochondria to Golgi. Typically, such comparisons display data clumps with weak correlations. **Bottom:** When fitted to repertoire equations by removing outliers (defined here as points not on or close to the line), the data clumps unfold into a set of parallel lines with R^2 close to 1.0. Since plots of control and experimental comparisons typically form clumps, this becomes a useful tool for extracting quantitative patterns from otherwise noisy data sets (From Bolender, 2004).

By applying this type of analysis to data pairs distributed throughout the biological hierarchy and coming from many different species, we can infer that quantitative relationships between and among parts exists as a universal property of living systems.

2-5 Ladder Equations

Equations allow us to begin the task of defining the fundamental structure of biology by form-

ing patterns that nest, connect, unfold and fold. To demonstrate such properties, we can reduce the 50,000+ ratio data (data pairs) in the universal biology database to a single ladder equation (Figure 2.5). It displays an exponential form with R^2 close to one ($Y = 0.000134e^{0.7498x}$, $R^2 = 0.999$). The Y intercept of this equation derives from the Y intercepts of 24 rung (power) equations (Bolender, 2003-2004). Although the algorithm used to make these calculations may or may not have a biological equivalent, it at least shows that order in biology can scale quantitatively by nesting equations. It suggests – at least theoretically - that biology could start with a single equation (rule) and apply unfolding algorithms to assemble a phenotype – having all its parts in the correct proportions. If biology stores such phenotypic templates in some real or virtual space, the curve at $t(1)$ of Figure 1.11 suggests where, when, and how we might begin to look for it mathematically.

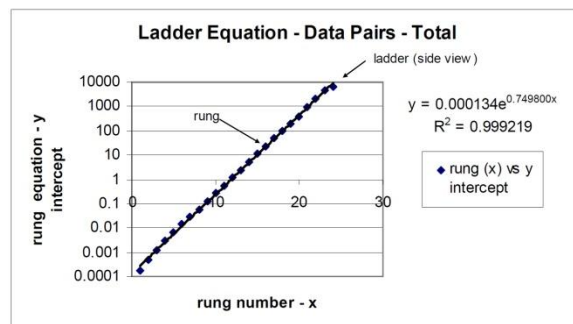


Figure 2.5 A ladder equation becomes the result of fitting a collection of 24 rung equation data (Y intercepts) to an exponential curve (From Bolender, 2003).

When we compare the ladder equation of Figure 2.5 to one derived from changes (Figure 2.6), the curves intersect and bear a striking resemblance to the solution of a linear programming problem. We would expect to find such a result if the structure of living things changes in such a way as to produce the optimal (best) outcome.

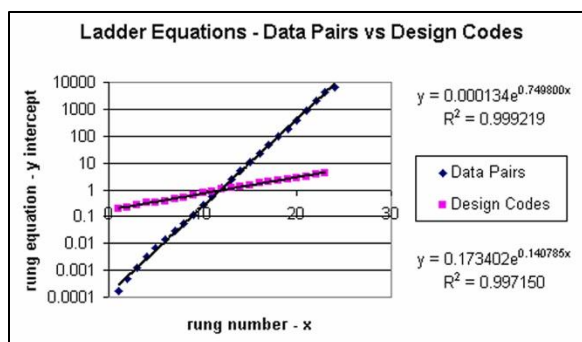


Figure 2.6 Two intersecting equations suggest that change in biology may represent an optimal solution (From Bolender, 2004).

2-6 Rung Equations

When plotted as power curves ($Y=bX^a$), rung equations display order as a set of 24 parallel regression curves, all having $R^2 \geq 0.999$. Figure 2.7 displays the equation for rung 15 of the ladder equation (Figure 2.5).

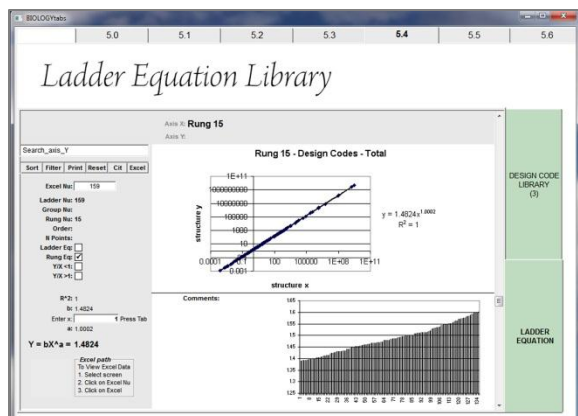


Figure 2.7 Rung equations represent regressions fitted to data pairs (From Bolender, 2004). Notice that the power curve ($Y=bX^a$) for rung 15 has a slope of 1.0002, which makes it effectively linear ($Y=bX$) (From Bolender, 2004).

Histograms of these rung equations provided an early indication that biological parts exhibit a stoichiometric order (Figure 2.8), expressed as ratios of whole numbers, and reminiscent of chemical valences. These biological valences can be seen in data pair tables by sorting on the ratio column where $X=1$ $Y=?$ (Figure 2.1).

This represents a new and important finding. Parts capable of existing in different valence

states, for example, must be contributing a substantial amount of noise to absolute data and concentrations because these measures provide only averages. As expected, this influence of the valence state on our understanding of change remains hidden until we express the data as ratios. We soon will discover that designing simulators for biology depends importantly on considering valences when predicting outcomes (Bolender, 2005-2011). By identifying valences as a distinct contributor to biological complexity, we have found another first principle of biology.

Figure 2.8 shows the relationship of mitochondria to other cell organelles - expressed as a ladder equation. As expected, the individual rung equations displayed a pattern reminiscent of the parallel regression lines (Figure 1.10 and Bolender, 2004). When plotted as a histogram, however, the ratios form distinct steps (Figure 2.8 - lower panel). This suggested that the relationship of one organelle to another describes a digital (discontinuous) rather than an analogue (continuous) distribution. This observation is particularly important because it will guide us to a new and more powerful digital format for our repertoire ratios and equations (Move 8).

The steps displayed in Figure 2.8 also suggest a deeper interpretation in that they may reflect the switching behavior of the genes responsible for producing the parts that form the ratios. Such a stepped pattern, for example, could be produced by similar sets of genes being turned on or off.

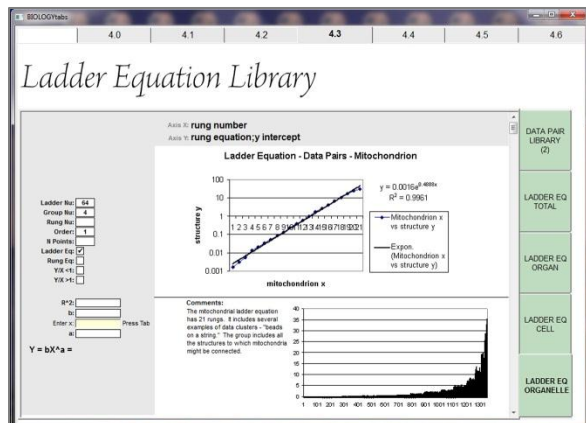


Figure 2.8 Ladder equation for mitochondria plotted against other cell organelles. Notice that when the data pair ratios (Y/X) are plotted as a histogram, steps appear (From Bolender, 2004).

Notice what happens when we plot ladder equations for all the organelles in the database. They display unique signatures. Each blue point in Figure 2.9 represents the y intercept of a power equation and each stack of points a different organelle view. The consistency in the linear separation of the blue points suggests the presence of discrete steps, as suggested by the histogram of Figure 2.8.

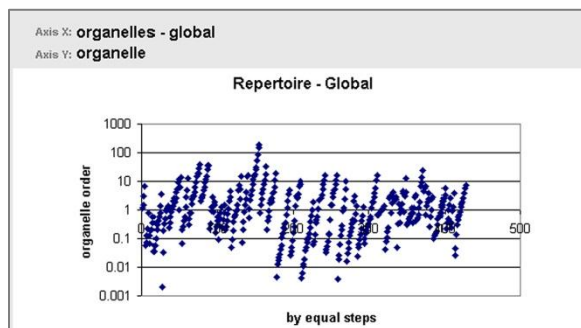


Figure 2.9 When the y intercepts of power equations are fitted to exponential equations, a global pattern of order can be seen (regression lines not drawn). Connections between organelles clearly appear to be ordered by rule (From Bolender, 2004).

Move 6: Can we extract patterns as equations from a table of data pair ratios?

Yes, data pair ratios serve as a rich source of patterns, captured as linear, power, and exponential equations. Equations, which often nest hieratically, extract order from clumps or clouds of data and reveal repeatedly the presence of step like patterns. Such patterns suggest rules based on a digital model, one that fits with the expected properties of valences and stoichiometry.

Since we now know how to extract patterns from published data as equations, we can use them to assemble simulators. Move 7 provides several examples.

Move 7: Can we assemble simulators with data pair ratios?

2-7 Simulators

When we connect a stack of repertoire equations, a change in the variable of one equation will spread to all the remaining equations. In effect, this predicts the outcomes for all the parts in a connected set. The program illustrated in Figure 2.10 includes a collection of simulators connecting the repertoire equations of eight organelles - nucleus, cytoplasm, lysosome, er, Golgi, lipid droplet, mitochondrion, and peroxisome. In the example shown, entering a new value for the nucleus and pressing the tab key triggers a cascade of changes to the other organelles. Notice that each organelle displays a column of new values, each of which carries a unique valence. Recall that same two parts can combine with different ratios of whole numbers. For example, mitochondria display 18 valences - globally.

Valences (expressed as ratios) play an essential role in defining a phenotype quantitatively. Since they can change in experimental settings, they add a key layer of detail to the diagnostic and predictive properties of a phenotype.

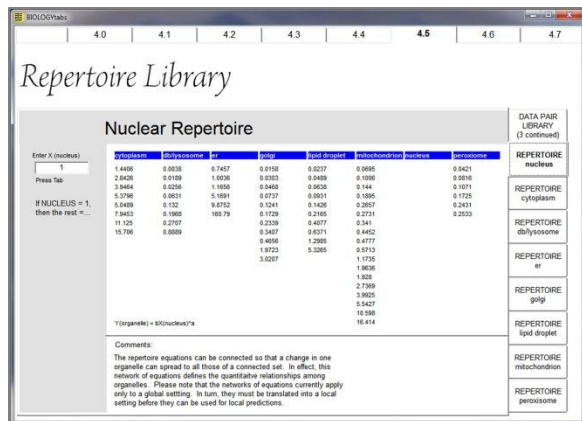


Figure 2.10 The nucleus displays several distinct relationships with other organelles. This simulator uses repertoire equations to connect changes in the nucleus to the organelles identified in the blue heading. Such data contribute to the finding that parts larger than molecules employ valences when making connections (From Bolender, 2004).

2-8 Reverse Engineering

If we take the hippocampus apart, for example, we can reconnect the parts with data pair ratios and reassemble the hippocampus with repertoire equations. In turn, entering a single seed value into the network of equations regenerates the original values or predicts new ones. By assigning a separate highlighting color to each animal type and applying it to each output value, we can compare the results of five animals (human, monkey mouse, rat, and shrew) across six parts of the hippocampus (Bolender, 2005).

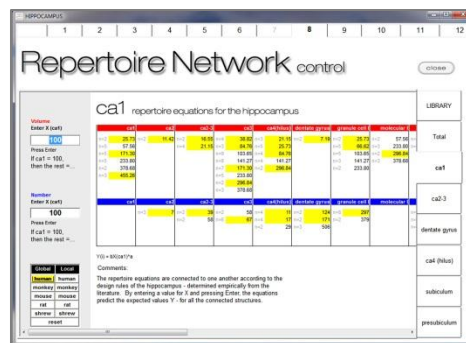


Figure 2.11 The hippocampus expressed as ratios of volumes (red) and cell numbers (blue). The yellow highlight identifies valences found in humans, which can be compared to those found in rats, mice, shrews, and monkeys. To predict the effects of a change in the hippocampus, enter a new value for ca1 and press Enter. The output shows what happens to biological parts when they change in a complexity (From Bolender, 2005).

These simulators demonstrate that we can reassemble the parts of organs post-mortem using data pair ratios and repertoire equations. More importantly, they uncover a previously unknown complexity in the data set as evidenced by the presence of multiple valences. By ignoring valences as a prominent feature of biological complexity, we diminish the information content of our data by accepting the lowered resolution of averages. Moreover, we may be distorting our statistical estimates of biological variation as well.

We need a quick reality check. The one thing we cannot know at this point in the game is the extent to which these valences are real or a result of our experimental methods. We can surmise, however, that both possibilities are most likely in play. Although we know that many biases exist, we continue to know little or nothing about the details of their magnitude or direction. This identifies a major weakness in our current experimental systems.

Move 7: Can we assemble simulators with data pair ratios?

Yes, provided we take into account the fact that ratios display valences. Valences effectively unfold the complexity of the relationship of one part to another as a set of well-defined patterns (ratios).

Thus far, the ratio data continue to suggest a digital rather than an analogue data structure for biology. In other words, outcomes often display a discontinuous (step like) distribution rather than a continuous one (Figures 2.3-2.11). The next move acts on this observation by converting the original data pair ratios to decimal ratios and decimal repertoire equations.

Move 8: Can we enhance our ability to detect patterns in data by shifting from an analogue to a digital format?

2-9 Decimal Ratios and Decimal Repertoire Equations

Patterns – derived from data pair ratios – uncovered previously undetectable relationships in the data of biological stereology (Moves 6 and 7; Bolender, 2004). When expressed as ratios (Y/X), however, the data pairs are somewhat unwieldy in that the absence of distinct boundaries between adjacent patterns creates challenges of interpretation when filtering and sorting the data tables.

This problem can be resolved by assigning the data pair ratios to distinct decimal bins, which are determined by calculating decimal repertoire equations for the data contained within each bin (Bolender, 2005) – a procedure analogous to the one described earlier for rung equations (2.6). This binning procedure condensed

more than 50,000 data pairs into just 81 decimal ratios accompanied by their decimal repertoire equations (Bolender, 2005).

The decimal steps were chosen such that the regressions predicted the published values with a maximum error of no more than $\pm 15\%$ - the typical error associated with stereological estimates (note that the original range (0.001 to 100,000) of the decimal steps was later revised (0.0001 to 100,000)). By switching to a decimal format, the data pairs displayed distinct boundaries, which allowed filtering and sorting routines to locate patterns quickly.

2-10 Extracting Hidden Information

By attaching each data pair ratio (Y/X) to a decimal repertoire equation, it became possible to extract hidden patterns – as ratios and equations – from large and otherwise amorphous clouds of data.

An illustrative example comes from the work of Seecharan et al. (2003) who counted cells in the lateral geniculate nucleus and retina, using 58 isogenic strains of mice. These authors were looking for neural connections in the visual system, but found only weak correlations (e.g., $R=0.44$; $R=0.33$) within the lateral geniculate nucleus and no correlations between the neurons in the nucleus and those in the retina.

After entering the cell counts from the lateral geniculate nucleus into the stereology database and forming data pairs, a pattern similar to the one reported in the original paper appeared (Figure 2.12; Top: Before). The R^2 of the regression line was 0.03.

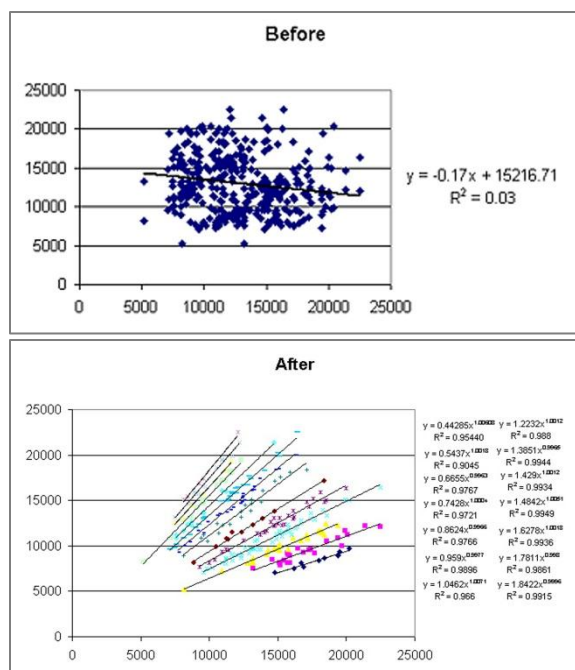


Figure 2.12 Top: Plotted cell counts from the lateral geniculate nucleus showed a single data cloud with little indication of order (After Seecharan et al. 2003; From Bolender, 2005). **Bottom:** The same amorphous data cloud was unfolded into 14 decimal repertoire equations with R^2 s >0.9. Notice how the decimal repertoire equations allow us to distinguish between closely related phenotypes quantitatively.

When analyzed with decimal repertoire equations, however, a very different picture emerged (Figure 2.12; Bottom: After). The data cloud was quickly resolved into 14 decimal repertoire equations with R^2 s >0.9.

The equations of Figure 2.12 tell us that in 58 isogenic strains of mice there are 14 ways to build a lateral geniculate nucleus. The table in Figure 2.13 summarizes these changes expressed as decimal ratios. Consider the three columns of the table highlighted in blue. Notice that each set of three ratios in a given row is a unique identifier of its specific strain. This concept of a unique identifier based on ratios will play a key role in Chapter 5 when we attempt a data driven approach to clinical diagnosis.

Nu	Cit Nu	Animal	Strain	Neu/Endo	Neu/Glia	Glia/Endo	Glia/Neu	Endo/Glia	Endo/Neu
53	4161	mouse	ce/j	0.3	0.6	0.6	1.6	1.5	2.5
44	4161	mouse	bxh12	0.3	0.5	0.6	1.7	1.5	2.6
20	4161	mouse	bxh23	0.4	0.8	0.5	1.1	1.9	2.2
2	4161	mouse	b6d2f1	0.4	0.7	0.6	1.3	1.5	2.0
30	4161	mouse	bxh34	0.4	0.6	0.7	1.5	1.3	2.1
43	4161	mouse	bxh11	0.4	1.0	0.8	1.7	1.1	2.0
42	4161	mouse	bxh10	0.4	0.6	0.7	1.5	1.3	2.1
23	4161	mouse	bxh27	0.4	0.6	0.7	1.5	1.3	2.0
57	4161	mouse	casa/rk	0.4	0.5	0.7	1.7	1.3	2.3
40	4161	mouse	bxh7	0.4	0.5	0.8	1.9	1.2	2.4
49	4161	mouse	balb/cbyj	0.4	0.6	0.6	1.5	1.4	2.2

Figure 2.13 A connection matrix illustrates the cell ratios in lateral geniculate nucleus of 58 isogenic strains of mice (Adapted from Seecharan et al. 2003; From Bolender, 2005). The cells include neurons (neu), glia, and endothelial cells (endo). Three cells taken two at a time give six data pair ratios. The blue highlight identifies a ratio <1, green=1, and red >1.

The unsettling finding of this study is that operating anywhere on the genome – adding or subtracting genes to produce isogenic strains – results in both intended and unintended consequences. In the isogenic strains, with genetic changes largely unrelated to the nervous system, the ratios of the cells (i.e., their valences) in the lateral geniculate nucleus showed extensive variability (Figures 2.12, 2.13). From this observation, one begins to suspect that in an organism, everything is connected and even small local perturbations can have global consequences. This suggests that genes may be playing the role of butterflies made famous by chaos theory (Kauffman, 1995; Walthrop, 1992). To wit, initial conditions can have enormous effects – a butterfly flapping its wings can trigger a storm at a location far removed.

Although modifying organisms genetically has become a routine operation, we cannot begin to know the consequences of such changes unless we begin to study biology as a complexity. The Seecharan data makes this point powerfully clear.

2-11 Growth Kinetics

Recall that the familiar growth curve for cells *in vitro* appears as an exponential curve, tailing off at both ends. In effect, biology expands a cell population by running a growth algorithm based on an exponential rule. Such a rule, however, suggests that the parts of these growing cells must be expanding exponentially as well.

Organelle data collected with stereological methods from keratinocytes in the epidermis (Klein-Szanto, 1977), expressed as ratios, and plotted as exponentials display the expected pattern in intact tissue.

As keratinocytes travel across the stratified squamous epithelium of the skin, their organelles grow according to an exponential rule. Connections between the endoplasmic reticulum (er) and associated organelles (mitochondria, melanosomes, lysosomes, lipid, and ribosomes) all fit exponentials. This would suggest that biology optimizes the growth of organelles in keratinocytes. Why? Recall that establishing a log growth phase for cells *in vitro* requires optimal growth conditions. By analogy, it would appear that biology can do the same *in vivo* (Bolender, 2005). Getting all the right parts in the right places at the right time to optimize an outcome suggests that keratinocytes know how to solve the extremely difficult problem of coordinating production, transport, and construction – perhaps perfectly. For us such a solution would be equivalent to solving a very tall stack of simultaneous equations. Given such an outcome, we now have another candidate for a first principle, namely optimization.

Move 8: Can we enhance our ability to detect patterns in data by shifting from an analogue to a digital format?

Yes, data pair ratios expressed as decimal steps support a highly efficient way of filtering and sorting data with the goal of finding patterns. The ratios in a given step (bin) can be expected to vary no more than $\pm 15\%$ from the original values with most falling within a much narrower range.

The effect of this move has been to shift our ratio data to a digital (stepped) format, which simplifies our task of finding and interpreting patterns. The examples given continue to offer evidence that mathematical order exists in data sets derived from biology. In the next move, we will attempt to extend this effort to the entire database.

Move 9: Is it possible to phenotype biology - quantitatively – and to express the results as a biological blueprint?

In this move, we will use a universal biology database containing decimal repertoire ratios – from a wide range of animal species - to construct a biological blueprint. In turn, we will use the blueprint to look for equations, patterns, and generalizations that can tell us something about the biological rules basic to phenotypes.

2-12 Biological Blueprint

By unfolding biological complexity into standardized units of information (ratios and equations), we can arrange published data into a composite blueprint. The blueprint can tell us how an organism defines its parts quantitatively and how a given phenotype retains or abandons

these relationships – within and across species. We will also revisit the issue of valences, wherein the same two parts can form different ratios.

When populating the blueprint, the data entry process consists of working through the list of data pair ratios in the universal biology database and recording the identity (X:Y) and frequency (n) of each ratio (Figure 2.14, Top). When completed, the blueprint summarizes the ratios, valences, and frequencies (Figure 2.14, Bottom).

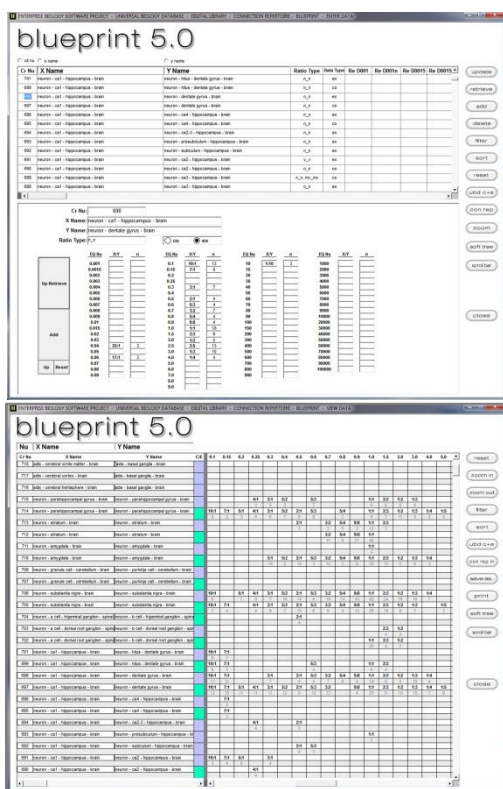
In effect, the blueprint offers an empirical overview of the mathematical core of biology – as it exists in the post-mortem data set of stereology. It shows that biological parts larger than molecules display a stoichiometry of whole numbers with many similarities existing within and across species. This suggests that organisms sharing the same parts with the same ratios are reading from the same playbook. In effect, the rules supervising these shared ratios appear to be highly conserved across biology.

Figure 2.14 Top: Data entry consists of tabulating 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).

Order in biology seems to be contagious. The ratio of one part to another depends on a vast number of subparts all of which find their origin in the genome. If parts are highly ordered in the phenotype, can we assume that this order projects back to the genotype? If yes, then it should be possible to use the order in the phenotype to identify a corresponding order in the genotype. By choosing the direction of information flow, we gain a substantial advantage. Going from a story – the phenotype - back to the words – the genes – is going to be far easier than going from the words to the stories that has taken biology a very long time to write. If we think of genes as a library of templates with switches that tell stories by creating phenotypes, reading such stories would seem to require little more than figuring out how to read phenotypes quantitatively all the way back to the genome. Such is the promise of stereological data.

The blueprint also serves as a convenient reference table for looking up phenotypes. A given pair of parts (X, Y) often display several distinct valences, characterized as multiples of whole numbers (X:Y). The ratio of mitochondria to peroxisomes, for example, can be 10:1, 20:1, and 33:1 – depending on the cell, animal, and experimental setting. Recall that such phenotypic information becomes essential when writing simulations, constructing networks of equations, or trying to detect changes. Moreover, the blueprint suggests that biology has evolved a universal parts inventory that it draws from when assembling species, growing, making repairs, and adjusting to the disease process.

One of the many challenges of operating within a complexity includes identifying patterns of parts that exist as distinct relationships, including one to many and many to many. Fortunately,



ly, the query by example (QBE) feature of relational databases offers a ready solution. The QBE interface shown in Figure 2.15 allows the user to assemble a query by selecting items from drop down list boxes. Clicking on the Query button sends the query - translated into the Structured Query Language (SQL) - to the database, which promptly returns a response. The screen shown in Figure 2.15 (Top), for example, wants to know all the data pairs that occur in the ratio of 1:2. The response includes 77 data pairs (Figure 2.14, Bottom).

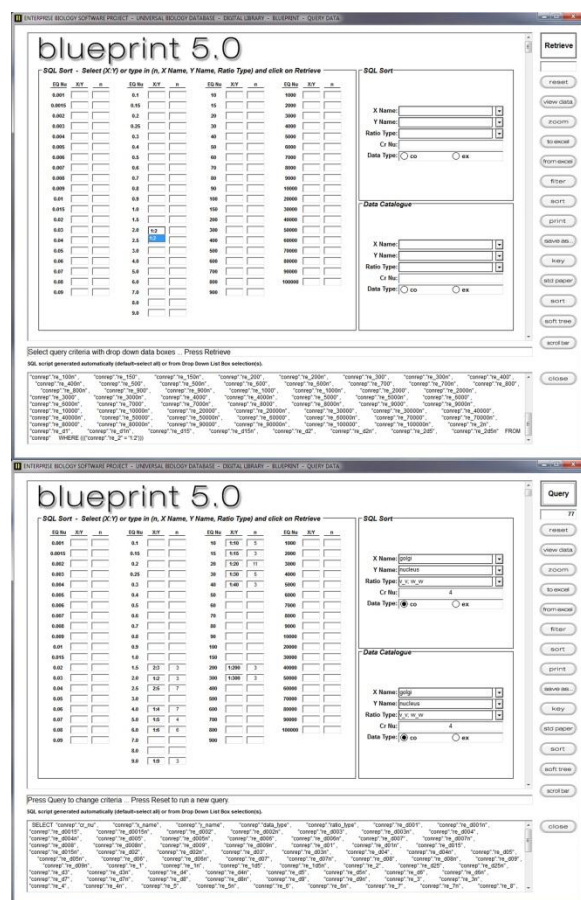


Figure 2.15 Top: The SQL interface shows the selection of the X:Y ratio of 1:2. As items are selected from the query screen, the SQL script at the bottom of the screen updates accordingly. **Bottom:** Clicking on the Query Button sends the request to the database, which promptly returns the information requested (From Bolender, 2006).

Finally, we can use the blueprint to look for generalizations. When expressed as a histogram (Figure 2.16), a summary of the entire blueprint table (Figure 2.14 bottom) shows that biology uses only about 50 decimal repertoire ratios, with far fewer doing most of the work (Figure 2.17; Bolender, 2006). Notice that the ratios identify five major peaks of activity.

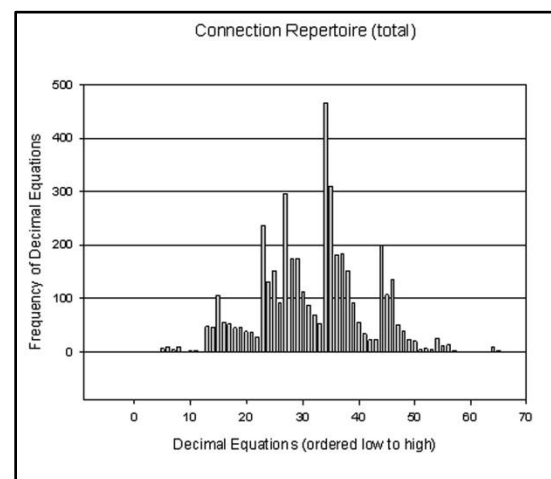


Figure 2.16 Distribution of the decimal repertoire ratios and corresponding equations in the biological blueprint (From Bolender, 2006).

Of the total blueprint entries (4,296) in Figure 2.16, roughly 40% include just six decimal repertoire equations (ratios): 50 to 1, 10 to 1, 3 to 1, 1 to 1, 2 to 3, and 1 to 10 (Figure 2.17).

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 2.17 The table lists the ratios in the biological blueprint occurring with the greatest frequency (From Bolender, 2006).

For neurons, the percentage of the most popular ratios goes up to about 70%. Neurons use six decimal repertoire equations defining five ratios: 3 to 1, 2 to 1, 3 to 2, 1 to 1, and 2 to 3.

The point to take from Figures 2.16 and 2.17 is that biology appears to be controlling the relationship of one part to another quite specifical-

ly. 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 rules. By simply forming ratios of small whole numbers, we can find the rules biology uses to order its parts. Stoichiometry, a first principle of chemistry, seems to be a first principle of biology as well.

In fact, first principles often generalize. Harmony in music, for example, occurs when two pitches vibrate at frequencies in small integer ratios. Especially pleasing ratios include 2:1, 2:3, and 3:4, two of which resonate with ratios just found in the human brain. Whether biology is subject to such universal principles by default or if they are selected for, remains, of course, an open question.

Move 9: Is it possible to phenotype biology - quantitatively – and to express the results as a biological blueprint?

Yes, data pair ratios represent one of the most effective ways of phenotyping an organism quantitatively. When interpreted globally, these same ratios combine to create patterns that help to explain the way biology structures itself by rule.

Once we know that biological order is quantifiable and accessible, we can explore new ways of finding and creating patterns of our own. In the next move, we will combine ratios of parts into equations to quantify phenotypes.

Move 10: Can a polynomial equation contain enough information to capture quantitatively the properties of a phenotype?

By expressing phenotypes as single equations, we can compare them visually and quickly iden-

tify patterns of change in complex data sets. This allows us to explore the potential role of data pair ratios in diagnosis and prediction.

2-13 Connection Phenotypes

The connection phenotype is a set of parts (data pairs), plotted as a frequency distribution, and fitted to a polynomial regression. As such, it represents a convenient way of expressing and interpreting a large data set visually (Bolender, 2008).

When calculating a connection phenotype, concentrations can work as well as absolute values provided control and experimental data sets remain separate. A change is detected by observing differences in patterns, not by dividing an experimental point by its control. Concentration data are limited in that a change can be detected, but it cannot be explained in terms of changes occurring in the individual parts making up the concentration. This requires absolute data. By eliminating the need to divide one concentration by another and minimizing the effects of bias and biological variation by forming ratios, connection phenotypes can extract valuable information from concentration data.

Since many publications still attempt to detect biological changes by dividing one concentration by another, the connection phenotype method of analysis improves the reliability of concentration data for detecting changes by comparing patterns instead. Since an emerging goal of investigative biology is to develop methods for mining new information from the millions of papers populating the literature, we will return to this topic often.

Biology blueprint expressed as connection phenotypes: Figure 2.18 translates the entire data set of the biological blueprint (Move 9) into two polynomial equations, one for control (blue) and the other for experimental (red).

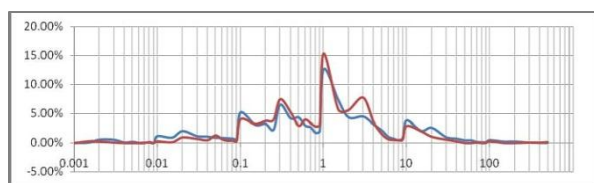


Figure 2.18 The connection phenotypes plot all the data pair ratios of the biological blueprint (CO=5224; EX=4095) as two polynomial equations (From Bolender, 2008). Blue identifies control and red experimental.

The plot shows how the data pair ratios of the controls can change in an experimental setting, wherein peaks and valleys alter their heights and locations, appear, or disappear. Such differences can be examined by identifying the data pair ratios at a given location along the x-axis - numbered from 0.001 to 500 (Figure 2.18). Methods for calculating connection phenotypes appear elsewhere (Bolender, 2008).

Growth in the rat adrenal: Patterns often undetected with standard methods of analysis become readily apparent when expressed as connection phenotypes. Figure 2.19, for example, shows 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 enlarges during development. The figure is of interest because it captures an ongoing growth program of biology with eight still frames (data columns). Such ratios may prove helpful in working out the orchestration of genetic events, which occur as an interaction between the genotype and the developing phenotype. Recall that we detected a similar pattern earlier with design code equations in Figure 1.11.

	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%
0.25	1	16.67%	0.00%	1	16.67%	0.00%	1	16.67%
0.3		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%	2
0.5	2	33.33%	0.00%	0.00%	1	16.67%	0.00%	0.00%
0.6	0.00%	0.00%	0.00%	1	16.67%	0.00%	0.00%	0.00%
0.7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1	16.67%
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%
1.5	2	33.33%	0.00%	1	16.67%	1	16.67%	0.00%
2	0.00%	2	33.33%	1	16.67%	1	16.67%	2
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%	1	16.67%
4	0.00%	1	16.67%	0.00%	1	16.67%	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 2.19 Development of the adrenal gland displays repeating patterns (yellow, green) over time (left to right) (Adapted from Nikicic et al., 1984; From Bolender, 2008).

ACTH and Growth in the hamster adrenal: Figure 2.20 shows the effect of ACTH on the development of the hamster adrenal. Once again, repeating patterns appear (yellow=yellow, green=green).

Days	0	7	14	21	28	35
DRV						
0.2	1	17%				
0.25	0%	1	17%		1	17%
0.3	0%	1	17%	2	33%	1
0.4	1	17%	0%	0%	0%	0%
0.5	1	17%	0%	0%	0%	0%
0.6	0%	0%	0%	0%	1	17%
0.7	0%	1	17%	0%	0%	1
0.8	0%	0%	0%	0%	0%	0%
0.9	0%	0%	0%	0%	0%	1
1	0%	1	17%	2	33%	1
1.5	1	17%	0%	0%	0%	0%
2	1	17%	0%	0%	0%	0%
2.5	0%	1	17%	0%	1	17%
3	0%	1	17%	2	33%	2
4	1	17%				

Figure 2.20 Notice that the response of the adrenal to ACTH during development also shows repeating patterns (yellow, green) over time (Adapted from Malendowicz, 1986; From Bolender, 2008).

Complexity of change: Connection phenotypes routinely uncover changes that absolute values miss. This can occur because comparing two data pair ratios involves the behavior of four variables, not just the two of a traditional control versus experimental comparison.

The top panel of the connect-the-dots display in Figure 2.21 joins each data pair (X Name: Y Name) found in the hippocampus of normal subjects to its counterpart in patients with Parkinson's disease. Although most of the connections suggest a change, the eight highlighted in

yellow suggest substantial differences in the range of 25 to 67% with a mean of 47%. In contrast, the original study reported no changes at all in the hippocampus (Harding et al., 2002).

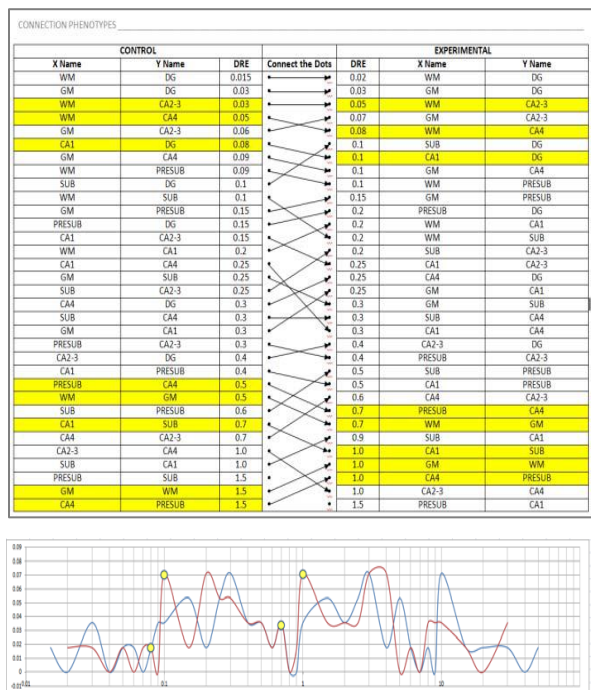


Figure 2.21 Top: In this connect the dots plot, the control hippocampus (left) is compared to one with Parkinson's Disease (right). The arrows identify the control and experimental locations of the same data pair ratio (Adapted from Harding et al., 2002; From Bolender, 2008). **Bottom:** Polynomial equations identify the two patient groups (normal=blue; Parkinson's=red).

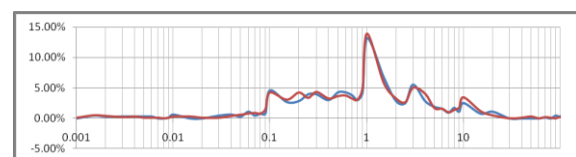
Figure 2.21 (Bottom) plots the data pairs as polynomial equations. The adjacent yellow dots identify the locations of two data pairs from the table highlighted in yellow. This allow us to see how specific data pairs move from the normal curve (blue) to the one representing patients with Parkinson's disease (red).

The value of the polynomial plots comes from their ability to provide the big picture without overwhelming the viewer with details. Such curves also encourage us to think about how diseases of the brain share similarities and differences and how these structural patterns

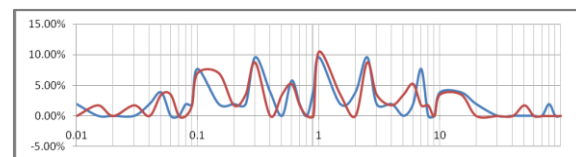
might relate to symptoms, diagnosis, and treatment protocols.

For example, we can readily compare plots of schizophrenia, epilepsy, Parkinson's disease, and alcoholism (Figure 2.22). Notice in the figure that the control (blue) and experimental (red) curves in three of the four diseases display surprisingly similar curves, even though each panel characterizes somewhat different parts of the brain. In contrast, epilepsy displays quite a different set of curves. Since the plots all derive from data collected post-mortem, their relationship to the patterns in living patients remains – for now - an open question.

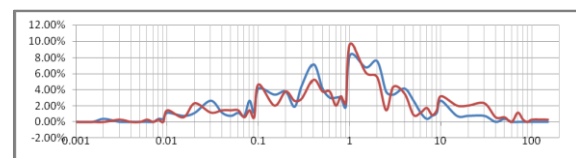
Schizophrenia – human



Epilepsy – human



Parkinson's disease – human



Alcoholism – human

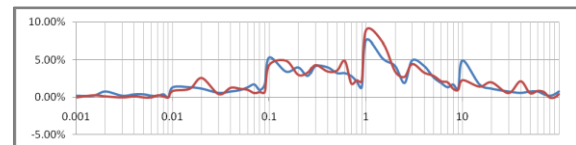


Figure 2.22 Connection phenotypes compare the polynomials of controls (blue) to experimentals (red) (From Bolender, 2008). The similarity of these patterns predicts a similarity in the design of these disorders.

In short, connection phenotypes can use tabular data or polynomial equations to identify quanti-

tative patterns in control and experimental data. They summarize large data sets, discover differences otherwise undetectable, and display diagnostic properties.

Move 10: Can a polynomial equation contain enough information to capture quantitatively the properties of a phenotype?

Yes, data pair ratios fitted to polynomial equations effectively capture the quantitative signature of a phenotype. By including concentration data, the approach increases - dramatically - our ability to identify changes.

Notice what is happening. Now that we are actively playing the complexity game with biology, we are learning that what we can do scientifically depends on the properties of our theory structure. By allowing our data to evolve from isolated data points to connected data pairs, we are now in the process of moving away from reductionism and toward complexity. In addition to detecting changes, our data can now produce equations, patterns, generalizations, and rules – all of which can uncover first principles.

To increase our level of play, we now have two options. We can increase the sample size or introduce more complexity. Alternatively, we can do both simultaneously by upgrading our playing field from data pairs to data triplets to data quadruplets.

Move 11: Can we increase the information content of our data by increasing the number of variables in the data ratio from two (X:Y) to three (X:Y:Z)?

2-14 Data Triplets

Using ratio data offers several advantages not the least of which is an 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 much more information.

Consider, for example, data triplets. We begin with three named parts A, B, and C and their respective values X, Y, and Z. Three parts taken two at a time, give six data pairs: A:B, A:C, B:A, B:C, C:A, C:B. After adding the values, we get AX:BY, AX:CZ, BY:AX, BY:CZ, CZ:AX, CZ:BY. A triplet exists when two data pairs share the same name and value. If BY=BY for data pairs AX:BY and BY:CZ, they form the triplet AX:BY:CZ. (Note: This manual approach to forming triplets will be replaced in the next chapter by one that operates automatically.) In turn, parts making up this triplet 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.

What is the point of including six copies of the same information, merely arranged in a different order? The short answer 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 in play. Moreover, if a single triplet misses a pattern because of the decimal binning of data, we still have several more chances to capture it.

Creating data ratios also has a profound multiplier effect, as shown in Figure 2.23. Notice that 5 original points taken two at a time yield 20 data pairs, 10 points yield 90, and 25 points yield 600. Taking N parts 3 at a time to form data triplets takes it up a notch - 5 points now yield 60 triplets, 10 yield 720, and 25 yield

Creating New Information

The graph illustrates the exponential growth of new information (ratios) as the number of original data points increases, for different tuple sizes. The Y-axis is logarithmic, representing the 'Number of New Ratios' from 1 to 1,000,000. The X-axis represents the 'Number of Original Data Points' from 0 to 30. Three series are shown: 2 (Data Pairs) in blue diamonds, 3 (Triplets) in red squares, and 4 (Quadruplets) in green triangles. All series show an upward trend, with 4 (Quadruplets) growing the fastest.

Number of Original Data Points	2 (Data Pairs)	3 (Triplets)	4 (Quadruplets)
5	20	60	200
10	100	700	5,000
15	200	3,500	35,000
20	400	7,000	120,000
25	700	12,000	350,000

2-15 Organism Codes

ORGANISM CODES

```
graph TD; A[BIOLOGY LITERATURE] --> B[STEREOLOGY LITERATURE DATABASE]; B --> C[UNIVERSAL BIOLOGY DATABASE]; C --> D[STEP 1 IDENTIFY AND MARK TRIPLETS]; D --> E[STEP 2 FORM TRIPLETS]; E --> F[STEP 3 CONNECT TRIPLETS]; F --> G[STEP 4 DRAW CONNECTIONS]; G --> H[STEP 5 STANDARDIZE DRAWINGS]; H --> I[Diagram 1: A network of 8 nodes (VAGUELY 0.0, GLOTTALLY 3.0, SILELY 0.0, PROCEEDING 0.0, LITHEOUS 0.0, METCH 3.0, LITHEUS 0.0, LITHEOUS 0.0) connected by lines]; I --> J[Diagram 2: A complex graph with 8 layers (LAYER V 0.0, LAYER III 0.0, LAYER VI 0.0, LAYER II 0.0, LAYER I 0.0, LAYER IV 0.0, LAYER VII 0.0, LAYER VIII 0.0) and a central node (LAYER V 0.0) connected to all other nodes];
```

The flowchart illustrates the process of creating organism codes. It begins with **BIOLOGY LITERATURE**, which is processed into a **STEREOLOGY LITERATURE DATABASE**, and then into a **UNIVERSAL BIOLOGY DATABASE**. The process then follows five steps: **STEP 1 IDENTIFY AND MARK TRIPLETS**, **STEP 2 FORM TRIPLETS**, **STEP 3 CONNECT TRIPLETS**, **STEP 4 DRAW CONNECTIONS**, and **STEP 5 STANDARDIZE DRAWINGS**. The final output is a complex graph structure, shown as two examples. The first example is a network of 8 nodes (VAGUELY 0.0, GLOTTALLY 3.0, SILELY 0.0, PROCEEDING 0.0, LITHEOUS 0.0, METCH 3.0, LITHEUS 0.0, LITHEOUS 0.0) connected by lines. The second example is a complex graph with 8 layers (LAYER V 0.0, LAYER III 0.0, LAYER VI 0.0, LAYER II 0.0, LAYER I 0.0, LAYER IV 0.0, LAYER VII 0.0, LAYER VIII 0.0) and a central node (LAYER V 0.0) connected to all other nodes.

An organism code shows how data are connected quantitatively (recall that we can combine two or more ratios when they share the same part and ratio). After generating more than 155 organism codes (Figure 2.25), it appears that a simple data pair represents only a very limited glimpse of a much larger connectivity (Bolender, 2010). Moreover, biology uses its connections to build in an extensive redundancy within its hierarchy of size. Apparently, once a set of design rules are in place, biology protects them by increasing the number of connections to a given part. Redundancy suggests yet another first principle.

45

stomach. This intermittent process requires a set of instructions – an algorithm - capable of moving large amounts of membrane from the cytoplasm to the cell surface and then triggering the release of HCl. By reversing the direction of the process, the membrane moves back into the cell and the secretion of HCl diminishes.

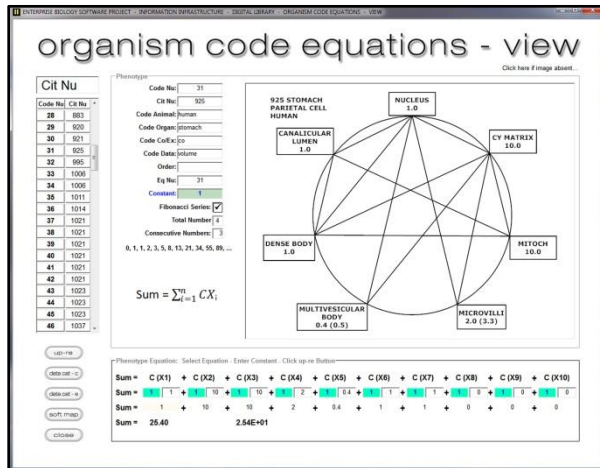


Figure 2.25 Organelles of the parietal cell in the human stomach display multiple connections. Moreover, the connections combine to form a string of ratios (lower panel), which presumably reflects the biological rule for constructing parietal cells - nuc(1) : cyma(10) : mito(10) : mivi(2) : mvb(0.4) : db(1) : calu(1) (Original data from Aase et al., 1976; From Bolender, 2010).

To biology, any change represents a complex event because it involves many parts and connections. Organism codes, which allow us to follow these changes graphically, allow us to watch biology behaving as a complexity.

Consider the next example shown in Figure 2.26. In going from health (normal) to disease (alcoholism, Alzheimer's), notice how the dentate gyrus of the hippocampus relinquishes its position as a central organizing structure to the presubiculum and how the parts and connections change – sometimes quite dramatically. In Alzheimer's disease, the normal redundancy of the connectivity disappears altogether.

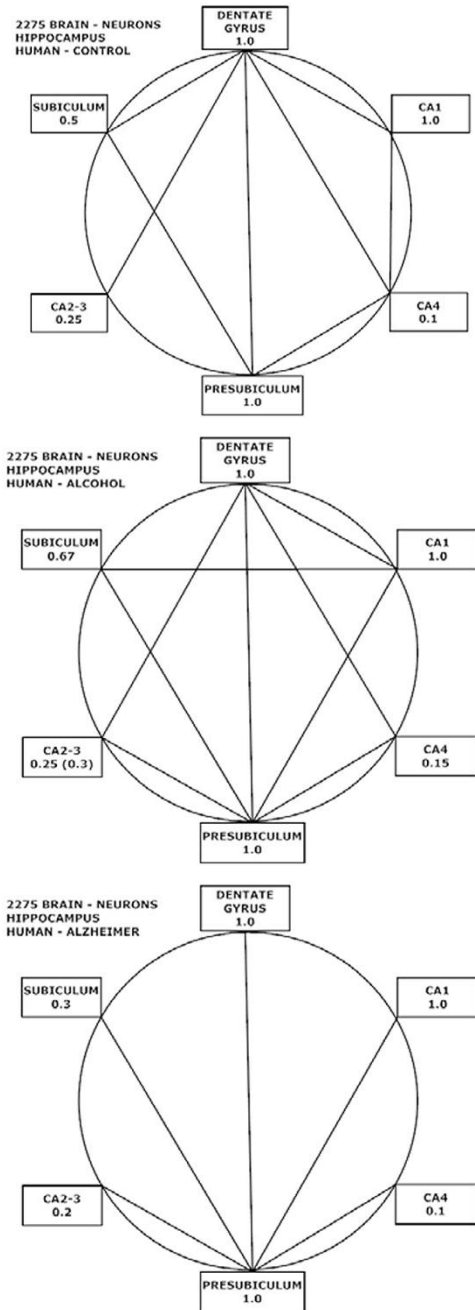


Figure 2.26 Organism codes - based on triplets - characterize the hippocampus in health (control: top) and disease (alcohol: middle, Alzheimer: bottom). Notice how triplets detect changes in the complexity of the hippocampus, using the relationship of parts to connections (Original data from Harding et al., 1997; From Bolender, 2010).

By scanning through the collection of organism codes (Figure 2.25), distinct patterns begin to

appear. Often, for example, the largest number of connections goes to a single part (node), which plays the role of a dominant central organizer. Searching the database for parts displaying this dominant behavior shows that the nucleus and mitochondrion assume this role most often (Figure 2.27). This pattern of dominance suggests a hierarchical control wherein cell parts tune their amounts most often with reference to nuclei and mitochondria. Moreover, the data also indicate that a tight linkage exists between the mitochondrial and nuclear compartments. One can imagine that such phenotypic patterns – or their antecedents – must be coded somewhere in the genome. Alternatively or additionally, self-organizing principles and feedback loops may be in play.

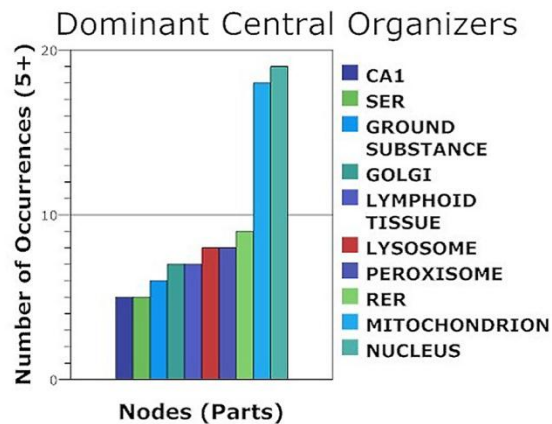


Figure 2.27 Dominant central organizers receive the largest number of connections (From Bolender, 2010). The tendency of cell organelles to key on the mitochondrion and nucleus may identify a first principle of cell design and dynamics.

Move 11: Can we increase the information content of our data by increasing the number of variables in the data ratio from two (X:Y) to three (X:Y:Z)?

Yes, data triplet ratios markedly increase the amount and complexity of our data. Organism codes summarize the data of a paper graphically as a collection of parts and connections that can undergo dramatic changes. Moreover, the connections reveal the existence of control centers that can shift in response to changing conditions.

2-16 Fibonacci Numbers

One of the best known patterns in nature is known as the Fibonacci series (0, 1, 1, 2, 3, 5, 8, ..., n), wherein adjacent numbers are added - in order – starting with [0,1] and continuing from left to right. This pattern of arranging parts occurs, for example, in DNA, flowers, vegetables, fingers, faces, and spiral galaxies. It also occurs in organism codes, most notably in experimental setting when the parts undergo changes (Bolender, 2010). Notice, once again, that the pattern relies on ratios of whole numbers.

Move 12: Is there more than one way to detect biological changes?

Throughout this chapter, we have changed our approach to detecting biological changes. Instead of dividing an experimental data point by its control, we compared the patterns produced by control and experimental ratios. This allowed us to avoid the concentration trap (Figure 1.7) because creating a ratio from either a

concentration or an absolute value gives the same result – provided the rules are obeyed (Figure 2.28). Even the seemingly amorphous clouds of data points we attribute to biological variation condense into highly ordered equations when we choose to interpret these same data points as ratios (e.g., Figure 2.12).

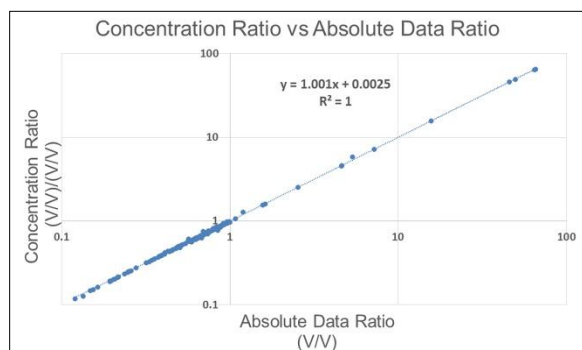


Figure 2.28 Forming a ratio of concentrations is equivalent to forming a ratio of absolute values, provided the concentrations share the same reference space. Using the concentration trap work screen (Figure 1.7), a representative sample of data points (~200) was taken, formed into ratios, and plotted. Notice the observed ($Y = 1.001x + 0.0025$) and expected ($Y = X$) equations are virtually identical.

Shifting from isolated points to ratios, however, is not without consequences. Statistical tests, for example, have close ties to the reductionist model and often depend importantly on distributions of isolated data points. Detecting a significant difference by comparing distributions involves a very different technology from the one needed to compare one pattern to another. A pattern either exists or not – period. Although biological variation can still exist within a ratio, it is being buffered by storing the data in decimal bins.

Move 12: Is there more than one way to detect biological changes?

Yes, in addition to detecting a change by comparing two mean values, we can look for changes in patterns derived from ratios.

2-17 Summary of Chapter 2

Chapter 2 introduces ratios as the primary data type in an emerging theory structure based on biological complexity. These ratios direct our attention toward discovering mathematical patterns created by large amounts of connected data and away from comparing single data points to look for significant differences. In effect, we are learning how to construct and interpret mathematical phenotypes.

A data pair defines a fundamental unit of biological complexity as a quantitative connection (ratio) existing between the values of two named parts. This union cannot be broken down into simpler components without losing the complex properties derived from the relationship. As a basic building block of complexity, data pairs can form strings, modules, and networks of data that capture and define complexity by rule.

Chapter 3

Game 3 – Creating a Parallel Complexity

In this the third game, we introduce the concept of a parallel complexity – a collection of ratios expressed as mathematical markers that will serve as a proxy for biology. To assure that this proxy is as close to the original biology as possible, we will rely exclusively on data coming from living subjects. Moreover, we will begin to understand why it takes a complexity to study a complexity.

Game 3 begins by replacing the post-mortem playing fields of game 2 with living ones. This requires a new database populated with MRI data (see, for example, Keller and Roberts, 2009) collected from the brains of living patients.

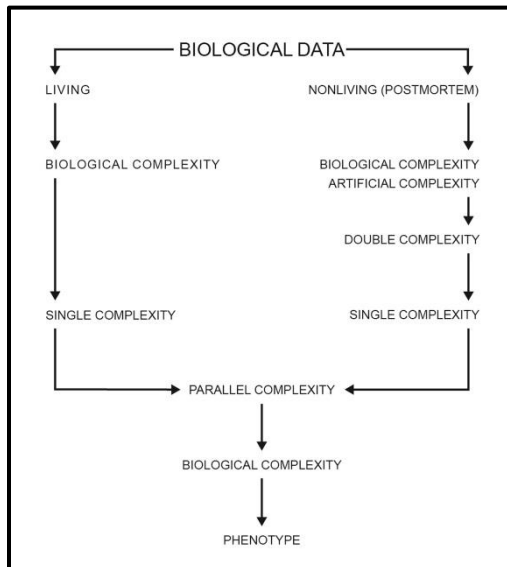


Figure 3.1 Extracting a phenotype from the biology literature requires a strategy driven by complexity theory.

To this end, we will begin with a paper that used MRI to estimate the volumes of 42 parts in the brains of normal and schizophrenic patients (Goldstein, et al., 1999). Since more than 97% of the data pairs formed triplets, the size of the

original data set increased from 42 parts to more than 2,000 ratios. To accommodate the size of this new playing field, we will turn to Mathematica (Wolfram Research, Inc.) for help with the calculations and graphics. We begin by mapping the parts of the brain mathematically.

Move 13: Can we map parts of the living human brain mathematically?

3-1 Mathematical Mapping

Mathematical mapping extends the reach of the organism codes described in section 2-15 by assembling a data set from all hierarchical levels. Moreover, by presenting the relationship of parts to connections visually, we can begin to grasp the magnitude of a complexity and become accustomed to viewing widespread changes in large patterns.

Figure 3.2 illustrates a mathematical map of the cerebral cortex of control subjects, wherein 42 parts (blue dots) display an astonishing number of connections (red lines). The figure begins to explain how biology uses its parts and connections to design the brain according to a set of well-defined rules (Bolender, 2011). By mapping all 42 parts simultaneously, we begin with a global view of the human cerebral cortex as a complexity, which we will then unfold progressively to discover local patterns and hidden relationships.

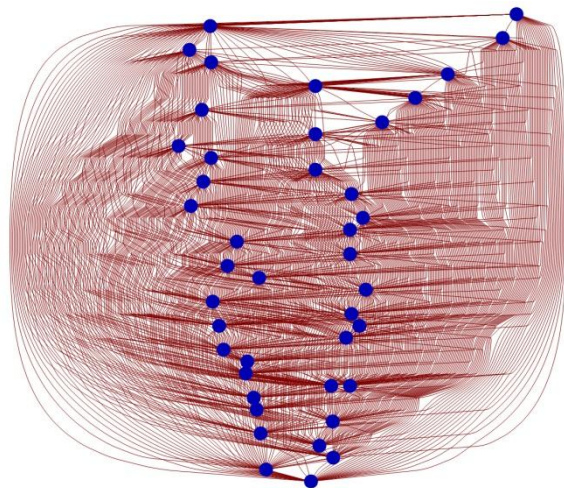


Figure 3.2 A mathematical map of the normal human cerebral cortex derives from forty two parts displaying thousands of connections (Adapted from Goldstein et al., 1999; From Bolender, 2011). If, for example, we change just one part, what happens to all the rest? Finding an answer to such a question requires a collaboration between biology and the biology literature. The parts (blue dots) include: angular gyrus, basal forebrain, central operculum, cerebral cortex, cingulate gyrus, cuneus, frontal lobe, frontal operculum, frontal pole, frontomedial cortex, frontoorbital cortex, fusiform gyrus oculo, fusiform gyrus telo, heschl gyrus, inferior frontal gyrus, inferior temporal gyrus, insula, lingual gyrus, medial paralimbic cortex, middle frontal gyrus, middle temporal gyrus, occipital lateral gyrus, occipital lobe, occipital pole, paracingulate cortex, parahippocampal gyrus, parietal lobe, parietal operculum, planum polare, planum temporal, postcentral gyrus, precentral frontal gyrus, precuneus, subcallosal cortex, superior frontal gyrus, superior parietal lobule, superior temporal gyrus, suppl motor cortex, supramarginal gyrus, telencephalon, temporal lobe, and temporal pole.

We begin the unfolding process by taking a closer look at the annular gyrus of the cerebral cortex. It can be isolated from Figure 3.2 and displayed with its parts (blue) and connections (red), as shown in Figure 3.3. Note that the topmost blue dot in the figure represents the angular gyrus. One is struck by the fact that a single biological part can be connected to so many other parts. In fact, all the original 42 parts are connected.

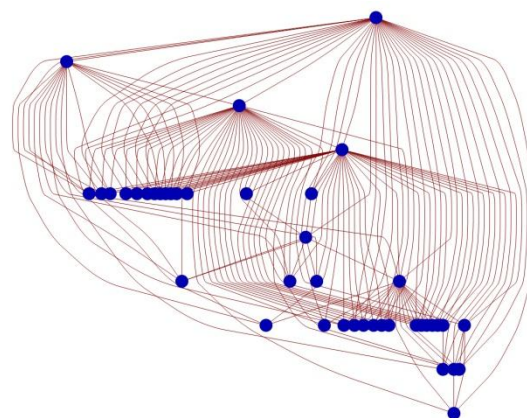


Figure 3.3 The mathematical map identifies the parts and connections of the angular gyrus in the normal human cerebral cortex. Note that all 42 parts are connected (Adapted from Goldstein et al., 1999; From Bolender, 2011).

Next, if we plot the two-dimensional map of Figure 3.3 in three dimensions (Figure 3.4), the underlying unit structure of the cortex appears as a collection of parts and connections forming triangles. Such triangular patterns often appear in biology (Bolender, 2010). Hagmann et al. (2008), for example, using physical mapping methods also found a similar triangular pattern in the human brain. This reappearing triangular pattern of connectivity suggests a modular design strategy may be in play.

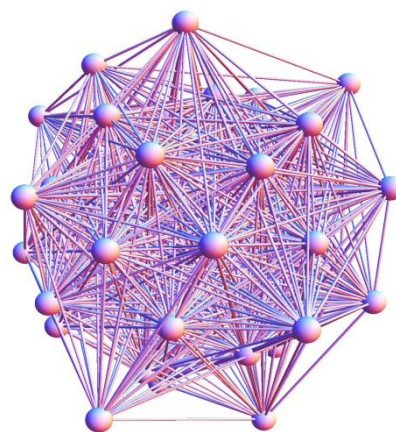


Figure 3.4 The 3D plot shows how 42 parts of the normal human cerebral cortex are interconnected (From Bolender, 2011).

From where do the connections come? Figure 3.5 shows the original 42 parts isolated according to the methods of reductionism. However, these parts still relate to one another by rule (stoichiometry), which, in this case, remains intact as volume ratios (Bolender, 2011). Since the parts define the ratios and the ratios the connections, we can recover a key element of complexity otherwise lost by our reductionist methods. Figure 3.5 becomes Figure 3.4 and then Figure 3.2.

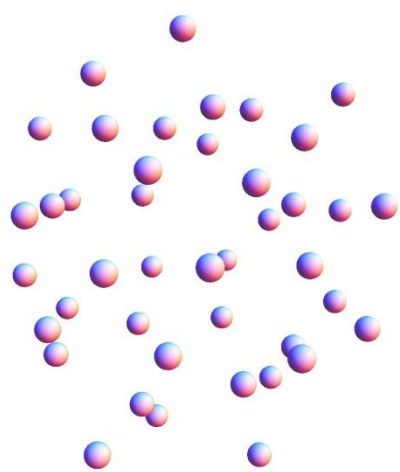


Figure 3.5 The original data set of the normal cerebral cortex included 42 isolated data points expressed as volumes (Adapted from Goldstein, et al., 1999; From Bolender, 2011).

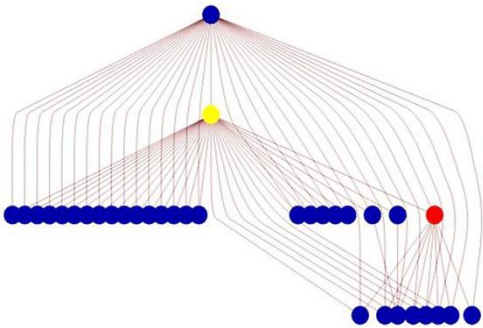
3-2 Diagnostic Patterns

Complexity offers a remarkable flexibility in the way we view biological change. Instead of looking for changes in just a few parts, we can now explore local and global patterns consisting of many parts and connections.

We can compare, for example, the frontal pole in normal individuals to those with schizophrenia (Figure 3.6). In the original study, no change was reported (Goldstein, et al., 1999). When viewed as a complexity, however, a distinct pattern of change becomes immediately apparent (Bolender, 2011). Schizophrenia produces dramatic changes in the connectivity of the parts

throughout the brain (Figures 3.6 – 3.8). In such cases, the images become diagnostic of the disorder.

Normal Patients



Patients with Schizophrenia

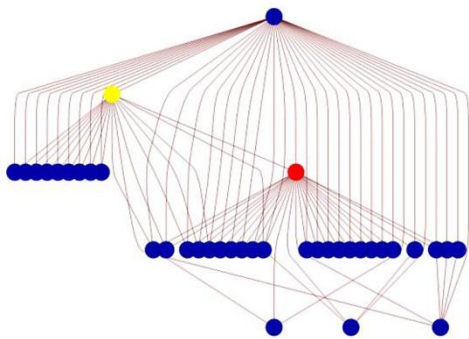


Figure 3.6 In schizophrenia, the relationship of parts to connections in the human frontal pole undergoes numerous changes (Adapted from Goldstein et al., 1999; From Bolender, 2011).

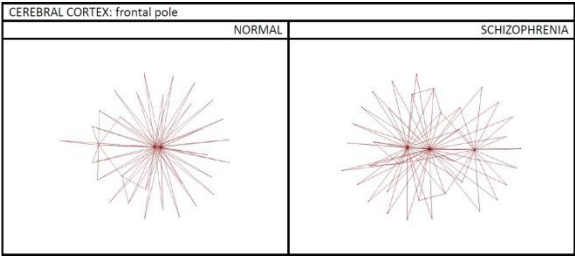


Figure 3.7 The frontal pole data of Figure 3.6 can also be plotted in three dimensions and rotated to view the changes in parts and connections (Adapted from Goldstein et al., 1999; From Bolender, 2011). By mapping - individually - all 42 parts of the cerebral cortex, we obtain a library of images for schizophrenia – one that offers a

comprehensive and perhaps more realistic view of this disorder as a complexity (see Appendix, Bolender, 2011).

Although complex patterns lend themselves to a graphical analysis, the results of mathematical mapping can also be expressed as equations (Figure 3.8). For example, the equation for schizophrenia (dotted red line) in 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 appears distinctly different from the corresponding control curve ($y = 0.0315x^2 - 3.4193x + 87.57$; $R^2 = 0.9854$).

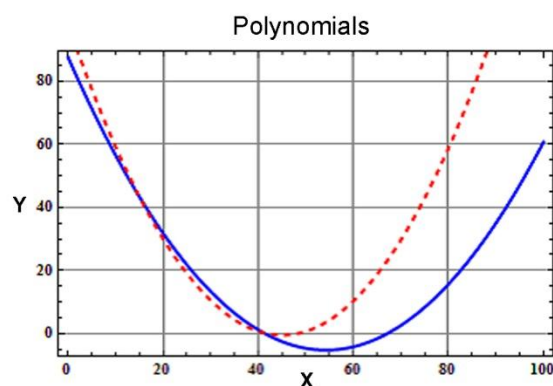


Figure 3.8 Diseases such as schizophrenia can be expressed as a polynomial equation (dashed red line), which can be distinguished from the one of normal patients (solid blue line) (From Bolender, 2011).

Move 13: Can we map parts of the living human brain mathematically?

Yes, the parts and connections of the brain map mathematically. Such mapping reveals widespread connectivity throughout the brain and suggests that diagnosing disorders mathematically may be possible – especially with MRI data from patients.

If we can map the complexity of a brain by using its parts and connections to characterize a phenotype, then it should also be possible to create

mathematical markers having diagnostic properties. This defines our next move.

Move 14: Can we diagnose disorders of the living human brain using mathematical markers?

3-3 Game Changer

At this point in the game, everything changes because we can now explore biology for the first time as a single complexity – a biology uncompromised by distortions of a post-mortem biology (Figure 3.1). This fortunate situation derives entirely from the online publication of the Internet Brain Volume Database (IBVD; Kennedy, et al., 2012; Poline et al., 2012). This splendid collection of published studies includes data from patients displaying a wide range of disorders - all of which translate into the universal data type of a parallel complexity – the mathematical marker.

3-4 Mathematical Markers

A mathematical marker has two components - the names of parts and their numerical values. Since we will be using triplets, the marker consists of a six character string (AX:BY:CZ) calculated as a ratio wherein $X=1$. It includes three named parts (A, B, C) with their corresponding numerical values (X, Y, Z). Markers capture the parts and connections of a phenotype locally and globally – within and across all levels of the biological hierarchy of size. For our purposes here, we will be using it as our basic unit of complexity.

Nested complexities, which exist throughout biology, can be managed effectively by translating dissimilar data sets into a single table of similar mathematical markers. This identifies a solution to the problem of integrating the wide range of existing biological data types. In effect, mathematical markers allow us to reconstruct

biology as a complexity using data from the biology literature.

Mathematical markers can be produced two ways, one way is hard the other easy. The hard way includes generating triplets individually by comparing data pairs (Figure 2.24). Although this approach works for relatively small data sets, it becomes impractical for large ones. Since we know empirically that roughly 84% of all the data pairs in our database form triplets (Bolender, 2012), it becomes easier just to transform all the data directly into triplets – one paper at a time.

This process includes the following steps. After entering the names of the parts from a given paper as a string (A, B, C ... N) into Mathematica (Wolfram Research), the program returns a list of all possible triplets - taking N parts three at a time. Next, the list of names is copied to an Excel spreadsheet where their numerical values are entered, used to calculate ratios, and assigned a decimal repertoire value. The final step consists of copying all the data from individual papers to a single spreadsheet, saving it as a text file, and importing the file into the table of a relational database.

When dealing with large data sets, a spreadsheet works best for data entry, whereas the database excels at finding complex patterns in large data sets. For further details of data entry, see Bolender (2012).

3-5 Diagnosing Disorders of the Brain (Shared Markers)

A clinical diagnosis identifies abnormalities in the phenotype, which - at any point in time - represents the current state of an individual. When, however, the symptoms and measures of one abnormality overlap those of others, a diagnosis depends on reconciling a combination of objective, subjective, and conflicting information. The unhappy consequence for the patient is that two or more well-qualified physi-

cians may view the same information and come to different conclusions. For brain disorders, getting to the correct diagnosis typically involves a lengthy and complicated process.

If instead, we treat the patient as a complexity, diagnosis becomes the product of a rule-based protocol. Mathematical markers provide this objective approach to diagnoses because we can design them to phenotype patients in health and disease. Since the IBVD provides access to the MRI data of at least 67 publications, we can generate more than 700,000 markers to capture the complexity of the phenotype in bewildering detail (Bolender, 2012).

We will discover in this chapter that diagnosing disorders of the brain objectively requires a large and representative database of standards, accompanied by a carefully crafted strategy for dealing with false positives and negatives (Chapter 5). To begin, we need to assemble a diagnostic tool for brain disorders, one that will allow us to run unknowns against known standards. A standard represents a set of markers known to be associated with a given disorder. For the examples included herein, the unknowns used for the test will come from publications not included in the diagnosis database of standards.

Performing a diagnosis consists of running the mathematical markers of an unknown disorder against a panel of standards (consisting of 24 known disorders) and tallying those that match (standard=unknown). The standard with the largest number of matches to the unknown markers is given the diagnosis (Figure 3.9). Although this approach worked successfully when applied to a small number of selected test cases, exhaustive testing was impractical because all the scoring was done manually (Bolender, 2012). In Chapter 5, we will automate this testing procedure, revisit the diagnosis problem, and test it rigorously.

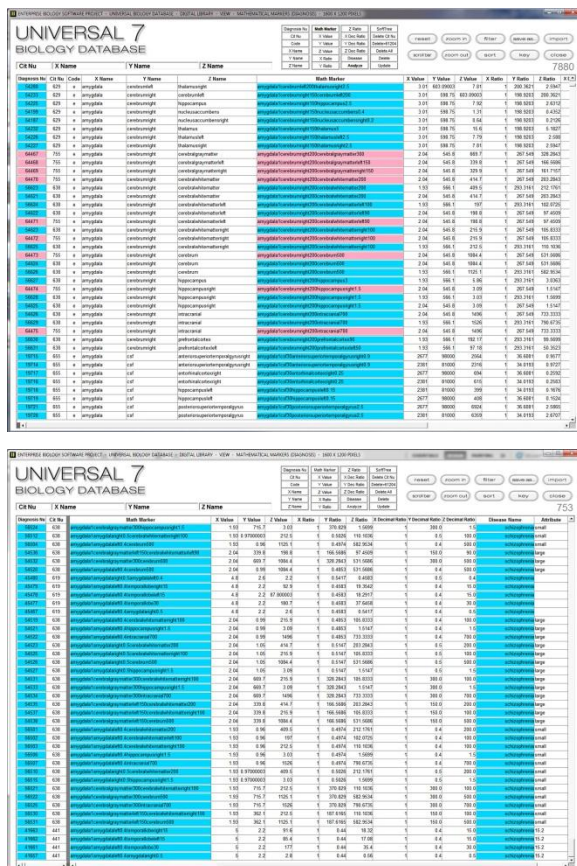


Figure 3.9 Mathematical markers use technology to diagnose unknown disorders of the brain objectively. Top: The diagnostic procedure consists of mixing the markers of 24 known disorders (blue) with markers coming from an unknown disorder (red) and checking off duplicates that occur between known and unknown markers. Bottom: After identifying all the duplicates, clicking on the analysis button summarizes the results. The disorder with the largest number of hits becomes the diagnosis (From Bolender, 2012).

Our immediate concern here will be to flesh out the characteristics of the diagnostic method by inspecting the patterns of individual disorders. Examples will include the bipolar disorder, Alzheimer's disease, and autism.

Bipolar Disorder: Mathematical markers provide new and sometimes surprising insights into the properties of brain disorders. When completely unfolded into a collection of markers, a given disorder shares many of its markers with those of other disorders. This pattern is seen in

a snapshot of a diagnostic table (Figure 3.10), summarized graphically as a histogram in Figure 3.11, and plotted as an equation in Figure 3.12. The unknown disorder was diagnosed correctly as bipolar (Figure 3.11).

1	6	0.4	unknown		X	
1	6	0.4	unknown		X	
1	6	2	unknown		X	
1	6	1	unknown		X	
1	6	1	unknown		X	
1	6	3	alzheimer	female		X
1	6	3	bipolar	type-one		X
1	6	3	unknown		X	X
1	6	3	alzheimer	female		X
1	6	3	bipolar	type-one		X
1	6	3	unknown		X	X
1	6	9	unknown		X	
1	6	4	alzheimer	female		X
1	6	4	bipolar	type-one		X
1	6	4	unknown		X	X
1	6	4	alzheimer	female		X
1	6	4	bipolar	type-one		X
1	6	4	unknown		X	X

Figure 3.10 Notice that the same marker (e.g., 1:6:3 and 1:6:4) can apply to more than one disease. Bipolar, alzheimer, and unknown all share the same mathematical marker (amygdalaright1putamen6putamenleft3) having the ratio 1:6:3 (From Bolender, 2012).

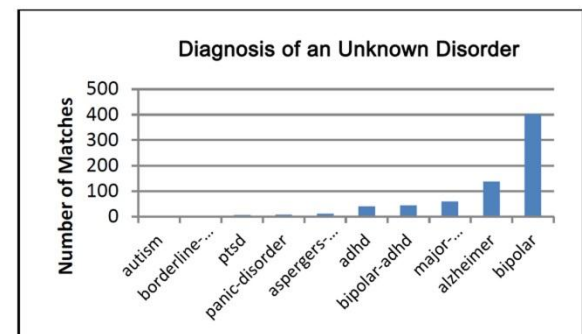


Figure 3.11 The histogram itemizes the disorders that share mathematical markers with the unknown, which was diagnosed correctly as bipolar. Bipolar disorder had 712 markers of which 401 (56.3%) were uniquely bipolar. The remaining 311 markers were shared with other disorders (From Bolender, 2012).

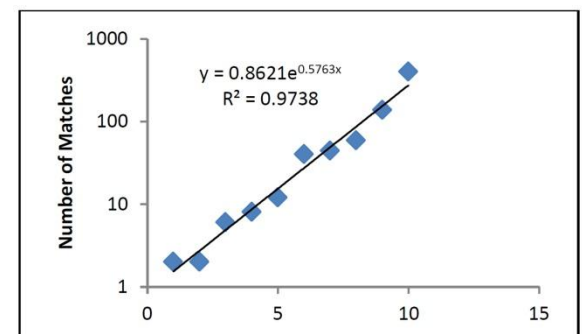


Figure 3.12 Disorders sharing mathematical markers with bipolar disorder fit an exponential equation, suggesting that mathematical markers might allow us to characterize a disorder with an equation. Since each disorder typically displays a range of values as it progresses, tracking this behavior with equations may offer a convenient diagnostic tool (From Bolender, 2012). Confidence in such equations will increase as the R^2 's approach 1.0.

Alzheimer's Disease: Notice that the mathematical markers of Alzheimer's disease (Figure 3.13) display a pattern of overlapping markers analogous to the one displayed in Figure 3.11. The disease, which carries 3,515 unique markers, shares 878 of its 4392 markers with 12 other disorders – from bipolar (606) to epilepsy (1). This pattern of sharing mathematical markers appears to occur throughout all brain disorders and may help to explain the difficulty encountered when using symptoms to make a diagnosis.

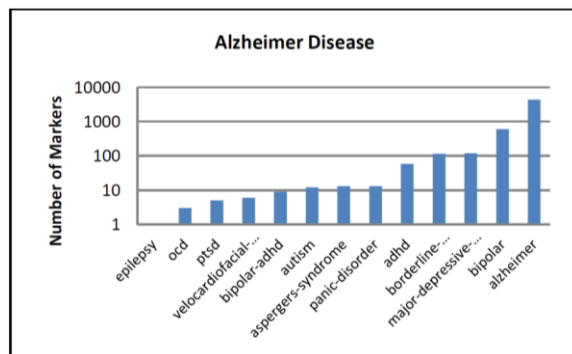


Figure 3.13 Many disorders share the same mathematical markers with Alzheimer's disease. Of the 4392 markers for Alzheimer's disease, 3514 (82%) are unique, whereas 606 (11%) are shared with bipolar disorder, 121 (2%) with major depressive disorder, 115 (2%) with borderline personality disorder, and 58 (1%) with ADHD. Note the logarithmic scale of the Y-axis (From Bolender, 2012).

Autism: Visualizing change in a complexity creates a challenge because it is so pervasive. Part of our job, therefore, becomes one of devising new ways of observing changes in patterns when large numbers of parts are involved – often numbering in the tens of thousands or millions.

Recall that when generating the ratios of mathematical markers ($AX:BY:CZ$), all the values are divided by X to set the value of X equal to 1.0. This allows us to plot mathematical markers as scatter plots with the two remaining variables (Y and Z). By preparing plots for normal and disease states and then flipping back and forth between the images, one discovers a spectacular amount of change. Entire clouds of points shift in and out. The massive change that occurs in a large data set appears at first surprising and then somewhat frightening in that it shows us how change actually operates in a complexity. Our current practice of following the behavior of only a few variables at a time would seem to miss the reality of the big picture all together.

Figure 3.14 illustrates autism as a complex change by superimposing control (yellow) and experimental (blue) scatterplots, wherein overlapping data points appear green. The log-log plot - used to spread out the data - reveals an extensive pattern of change associated with autism. Notice that in this comparison of normal patients (yellow) to those with autism (blue), the blue points (autism) can move inward, outward, or stay the same. When the marker remains unchanged, it appears green (yellow + blue).

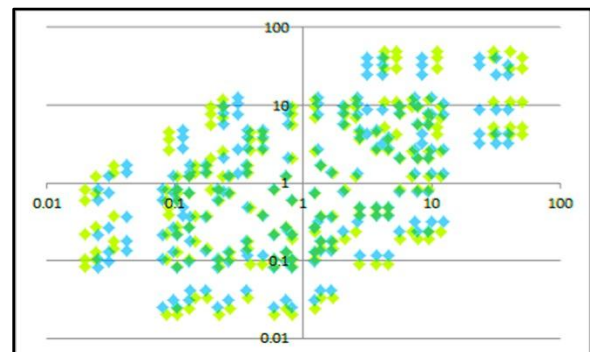


Figure 3.14 Key: Autism (blue), Normal (yellow), overlap (green). Compared to the normal, autism is characterized largely by a contraction of the point set (inward movement – yellow to blue). However, examples of an outer movement and of no movement (complete green squares - overlap of blue and yellow) also appear in an animation.

Although mathematical markers offer an objective approach to diagnosis, we have not yet recognized or dealt with the problem of false positives and negatives. This will be done in Chapter 5. We did learn, however, that a considerable overlap exists between the markers of different disorders. This finding takes us one-step closer to understanding the disease process.

Move 14: Can we diagnose disorders of the living human brain using mathematical markers?

Yes, mathematical markers derived from published MRI data offer a new strategy for diagnosing disorders of the brain. The approach, however, still requires comprehensive testing to ensure its reliability.

Since our current MRI playing field provides a rich collection of both local and global patterns, we can begin to look for generalizations. In the next move, we will start this process by unfolding phenotypes into mathematical markers to get a closer look at the way the brain assembles disorders.

Move 15: Do disorders of the brain derive from a common design plan?

Using mathematical markers derived from MRI, we can isolate, analyze, and compare the complexity of disorders one on one or as an entire group. We can do this because living systems

display a remarkable property - the same patterns seen locally also appear globally.

Although forming ratios appears to minimize the effects of bias and animal variation, the power of MRI data derives from the fact that most of its complexity comes from biology. Were this not the case, then we would not expect to see such widespread agreement between local and global markers. This tells us that data collected from living individuals can serve as a gold standard to which all other data types can be compared.

In short, creating our parallel complexities with MRI data simplifies our job enormously because dealing with one source of complexity is much easier than dealing with two (Figure 3.1).

3-6 Generalizing Disorders

We know that biology defines itself as a complexity, one that we can unfold into connected parts using mathematical markers. In turn, the order given to our parallel complexity by these markers advances our level of play to that of pursuing generalizations related to the disease process.

Each disorder of the human brain carries a distinctive set of markers. However, as shown in Figures 3.11 and 3.13, the same marker can appear in more than one disorder. This means that the complexity of a disease depends on several factors, including the composition of individual markers, the presence or absence of specific markers, and the total number of markers in play. By unfolding each disorder into a collection of well-ordered and clearly identified markers, we can define it as a unique phenotype.

A table of parts cross-correlated with disorders begins the process of generalizing the design strategy exercised by biology in the brain (Figure 3.15). Notice that specific parts define a disorder, that different disorders often share the same parts, and that a relatively small num-

ber of parts (35/185 or 19%) accounts for most of the disorder. Schizophrenia (26 parts) and bipolar disorder (20 parts) seem to create the most damage, whereas the amygdala (13), caudate (13), hippocampus (10), putamen (10), and thalamus (9) appear most vulnerable.

	adhd	alcohol	alzheimer	aspergers	autism	bipolar	borderline_per	dev_delayed	down	dyslexia	epilepsy	fragilex	huntington	kleinfelter	major_depress	ocd	panic_disorder	pttd	schizophrenia	velocardiac	williams
amygdala																					
anteriorsuperiortemporalgyrus																					
basalganglia																					
brain																					
caudate																					
cerebellargraymatter																					
cerebellum																					
cerebralcortex																					
cerebralgaymatter																					
cerebralwhitematter																					
cerebrum																					
cst																					
entorhinalcortex																					
globuspallidus																					
graymatter																					
hippocampalcomplex																					
hippocampus																					
hypothalamus																					
insula																					
intracranial																					
lateralventricle																					
nucleusaccumbens																					
occipitallobe																					
pallidum																					
parahippocampalcomplex																					
parietallobe																					
posteriorinsula																					
posteriorsuperiortemporalgyrus																					
prefrontalcortex																					
putamen																					
striatum																					
temporallobe																					
thalamus																					
thirdventricle																					
ventraldiencephalon																					
whitematter																					

Figure 3.15 The table summarizes the involvement of specific parts in 21 different disorders of the human brain. 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 data come from the diagnosis database (From Bolender, 2012).

3-7 Playing the Disorder Game

Figure 3-15 suggests that biology assembles disorders by mixing and matching parts from a common pool of resources. As a complexity, living things seem to be the product of a nature that likes to build things from well-defined sets of parts, be they normal or abnormal.

By expressing the relationship of one disorder to another graphically, we can begin to tease out some of the details of this modular strategy.

Consider schizophrenia. It represents the most extensive departure from the norm in that it carries at least 123 abnormal markers (Figure 3.16).

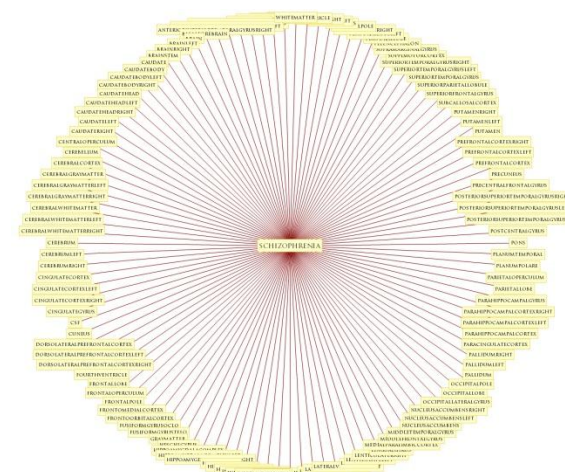


Figure 3.16 Schizophrenia results from the disruption of at least 123 parts of the brain (From Bolender, 2012). Enlarge as needed or view the originals on the internet (enterprisebiology.com).

Notice what happens when we add 14 other disorders to the plot of Figure 3.16. Although schizophrenia remains dominant with its 123 parts and connections, it shares many of its parts (~30%) with other disorders (Figure 3.17). One unexpected finding is that the parts and connections of six individually recognizable disorders are identical to those of schizophrenia. What might this mean? If a program for schizophrenia exists, is it read only partially to produce one of these six disorders or is schizophrenia the accumulation of many different disorders? Such questions, of course, go to the heart of the disease process.

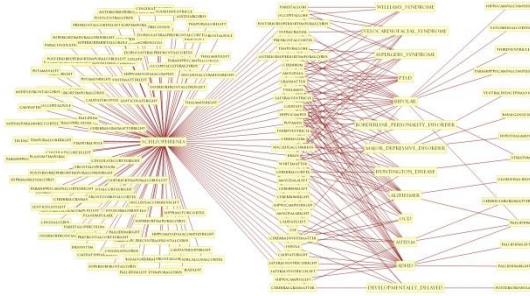


Figure 3.17 Disorders of the brain share many similar parts and connections (From Bolender, 2012). Enlarge the image to view details.

When we plot just schizophrenia, bipolar disorder, and ADHD, the complex relationship of one disorder to another becomes more apparent. Bipolar disorder and ADHD occur as a distinct subset of schizophrenia in that they share 80% of the same parts and connections. Moreover, a relationship exists between bipolar disorder and ADHD in that they share roughly 25% of the same parts and connections.

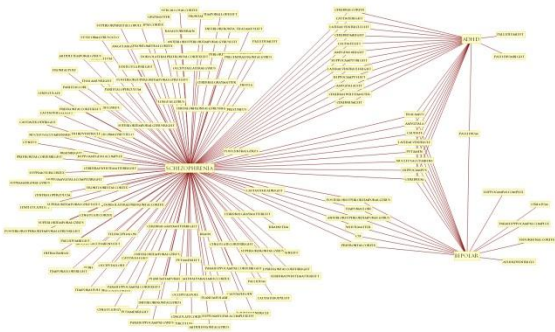


Figure 3.18 ADHD and bipolar disorder share many identical parts and connections with schizophrenia, as well as with each other (From Bolender, 2012). Enlarge as needed.

This pattern of a close relationship between disorders (Figure 3.18) persists as a general pattern. Figure 3.19, for example, shows the relationship of bipolar disorder to Alzheimer disease. They share 9 of 25 (36%) parts and connections.

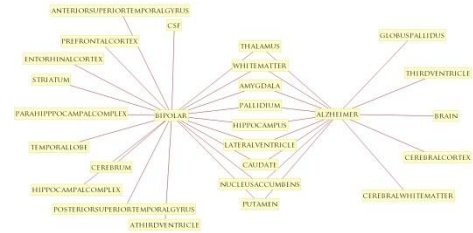


Figure 3.19 Bipolar disorder and Alzheimer disease share similar parts and connections (From Bolender, 2012).

This generalization also exists quantitatively (Chapter 6). Different disorders of the brain share not only parts and connections, but also identical mathematical markers. In effect, evidence that biology constructs normal and abnormal brains as complexities based on well-defined relationships of parts to connections continues to accumulate.

Apparently, biology is playing complexity games with its parts and connections to produce different brains with different properties. Sometimes this strategy gives us normal individuals, other times we get geniuses, savants, and great artists. Other times it produces disorders with harmful consequences. Mathematical markers tell us that all brains share a common design strategy, but that specific disorders resemble recipes in that they consist of well-defined populations of shared and unique components.

Given what we have learned thus far, studying individual disorders within the context of all disorders may 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 a host of other disorders. As we discovered earlier with the lateral geniculate nucleus (Figure 2.13), even a small change in the genome can trigger global consequences.

Move 15: Do disorders of the brain derive from a common design?

Yes, mathematical markers, which unfold the brain into basic units of complexity, allow us to generalize both the design and interrelationships of many brain disorders. The common threads found to be running through disorders begin to challenge our current perceptions of disease, diagnosis, treatment, and prevention.

3-8 Summary of Chapter 3

In chapter two, we explored ways of looking at biology as a complexity by folding and unfolding biological parts and connections mathematically. The resulting patterns told us that biology orders itself by creating and maintaining proportions of one part to another. This offered us assurance that biology exists as a rule-based system, one that we could read and interpret mathematically.

The central point to emerge from Chapter 3 is that it takes a complexity to understand a com-

plexity. This basic principle translates into a parallel complexity, which becomes a playing field designed according to rules consistent with those of biology. Since biology runs a considerable portion of its complexity business with ratios, we profit by running our parallel complexity business in exactly the same way.

By shifting to the MRI data of living subjects, new playing fields allowed us to map portions of the human brain mathematically and to begin the process of figuring out how to diagnose disorders of the brain objectively.

An especially important finding to emerge from the MRI data was that patterns can become global and lead directly to biological rules and generalizations. By learning how to read biology mathematically, different investigators interacting with different patients in different settings can now expect to find similar patterns when the same rules are in play. In other words, diagnosis becomes an exercise in accessing a phenotype objectively.

In chapter 4, we start a new game based on what we have learned thus far. The point of the game will be to explore the relationship of theory structure to scientific results. It will allow us to run a reality check on our current research model.

Chapter 4

Game 4 – Reconciling Differences

As a complexity, biology displays the expected patterns and behaviors of a rule based system. However, there is a problem. We remain committed to the widely held view that both living and nonliving states of biology yield similar data and information. If untrue, then interpreting research based on post-mortem material becomes problematic – especially for those studies dependent on detecting biological changes. This puts us in the difficult position of having to play a far riskier game in Chapter 4 – one with potentially far-reaching consequences.

Implicit in reductionist theory when applied to biology is the assumption that data coming from parts represent valid measures of biology, even when sampled post-mortem. Since we now have access to data coming from the same parts in living (IVBD) and post-mortem (Stereology Literature Database) brains, we will use this chapter to test this assumption and attempt to reconcile any differences we find.

Move 16: Can post-mortem data diagnose a disorder of the brain (schizophrenia) correctly – using mathematical markers?

4-1 Diagnosing Disorders Post-mortem

Given the assumption of data compatibility stated above, our ability to diagnose disorders in the brains should extend to both living and post-mortem brains. If not, then reductionist theory will have failed to withstand the challenge instigated by the move.

The first test consists of running post-mortem markers for schizophrenia (unknowns) against a comparable set of markers taken from living patients (knowns). Figure 4.1 indicates that the post-mortem data clearly missed the correct diagnosis of

schizophrenia, giving it instead to the bipolar disorder. Consequently, the results do not support the assumption of data equivalence. Moreover, the ensuing tests all lead to the same disappointing conclusion.

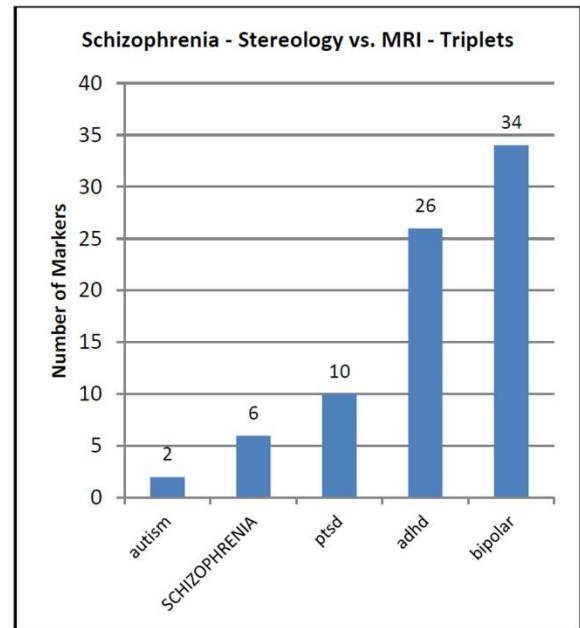


Figure 4.1 When unknown mathematical markers taken from post-mortem brains (stereology) were run against those of living brains (MRI), the resulting diagnosis (bipolar) was incorrect. The correct diagnosis - schizophrenia – was not even close. Apparently, living and non-living brains are very different quantitatively. Notice that relatively few markers were in play (From Bolender, 2013).

The next example attempted to make the test easier to pass by reducing the triplet markers (AX:BY:CZ) to data pairs (AX:BY). This gave a better outcome, but the results were too close to call. Figure 4.2 indicates that the diagnosis went to both schizophrenia and Alzheimer's disease with bipolar running a close second. In effect, the unavoidable conclusion to come from the tests thus far is that post-mortem brains no longer have many of the

quantitative patterns found in living brains (Bolender, 2012).

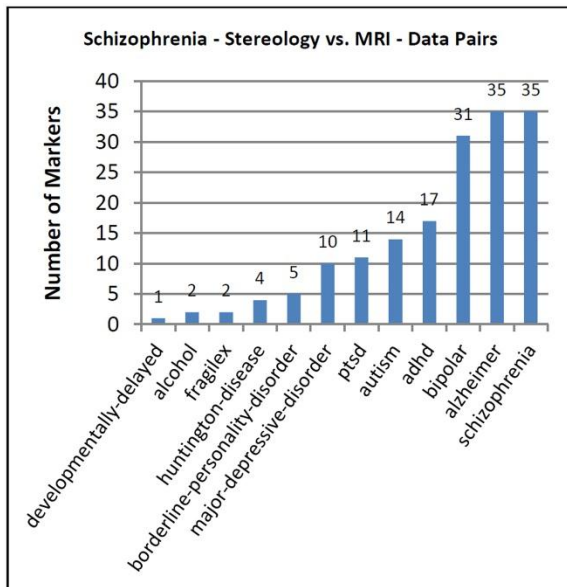


Figure 4.2 Using mathematical markers based on data pairs, post-mortem data still could not diagnose the unknown as schizophrenia. Schizophrenia tied with Alzheimer disease (From Bolender, 2013).

Had the diagnosis been successful, what pattern should have appeared in Figure 4.2? We would expect the histogram to resemble the distribution of data pair markers displayed by the living brain (Figure 4.3).

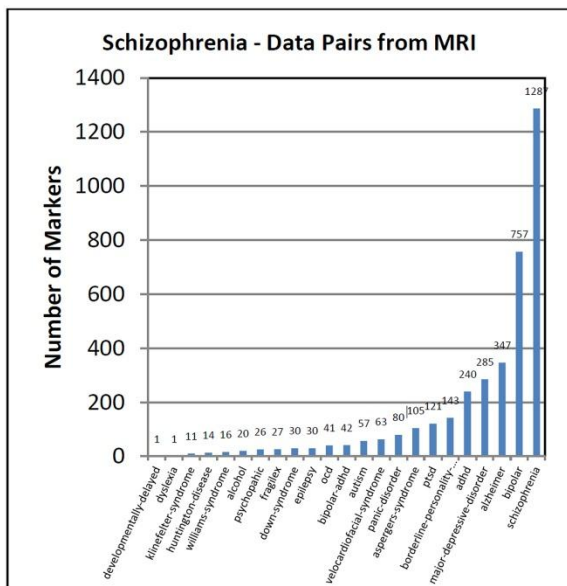


Figure 4.3 The MRI data of living brains provided 1287 mathematical markers for schizophrenia – based on data pairs (AX:BY). Notice that of these 1287 markers, 757 were shared with bipolar, 347 with Alzheimer, and 285 with major depressive disorder (From Bolender, 2013).

Move 16: Can post-mortem data diagnose a disorder of the brain (schizophrenia) correctly – using mathematical markers?

No, this was the wrong move. Markers from living brains cannot diagnose a disorder when run against unknown markers derived from post-mortem brains. The two data sets are largely incompatible.

If living and post-mortem brains show little compatibility between their mathematical markers, then extensive and unequal changes must have occurred to the volumes of the parts post-mortem. Such a finding puts at risk our assumption that we can gather biologically relevant information – as volumes or as data related to volumes – from post-mortem material. The next move will attempt to determine what happens to biological parts post-mortem.

Move 17: Can we determine the extent to which the parts of post-mortem brains differ from those of living brains?

4-2 Global Patterns in Normal Brains

When the same patterns occur across publications, they qualify as being global. Counts of duplicate (shared) mathematical markers in a database table – derived from different publications – serve to identify and measure of the persistence of a given global pattern.

By counting duplicate markers in control data sets coming from living and post-mortem brains separately, we can estimate the amount of information lost post-mortem. Markers are identified as duplicates when they appear in at least two publications, but for our purposes here, a given duplicate marker

is counted only once - even if it has many more copies.

Figure 4.4 shows that in making the transition from living to post-mortem, normal human brains experience a substantial loss in global properties. The analysis included 160,736 markers for MRI, of which 17,048 (10.6%) were duplicated in more than one paper. In contrast, the post-mortem data of stereology had 53,298 markers, of which 1,060 (2%) were duplicates. This tells us that the expected consequence of collecting data from post-mortem brains is an 81% loss in the information we would need to study the brain as a complexity.

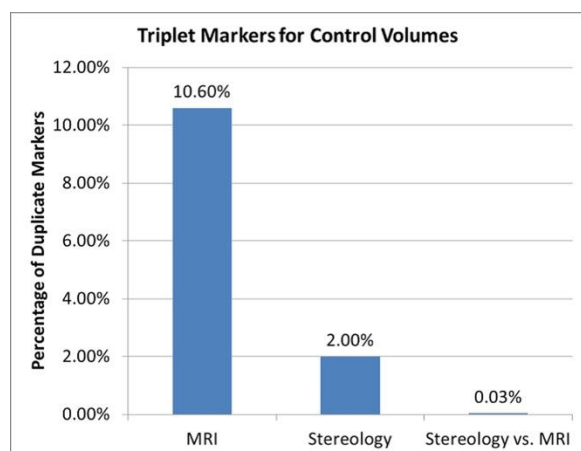


Figure 4.4 In the MRI database, living brains – across publications - displayed a global pattern as suggested by the almost 11% incidence of duplicate mathematical markers. In contrast, this measure in post-mortem brains of the stereology database was only 2%. This difference suggests an 81% loss of information - $((1-(2\%/10.6\%)) \times 100\% = 81\%)$ (From Bolender, 2013).

One of the most striking observations to come from living brains is that a given mathematical marker can appear routinely in many different publications (Bolender, 2012). However, Figure 4.4 suggests that this global property of markers in living brains all but disappears in post-mortem brains (Figure 4.4).

4-3 Disrupted Global Patterns in the Brain

Schizophrenia disrupts the patterns of a normal brain, as illustrated in Figure 3.16. We can evaluate the extent of this disruption – in living and post-

mortem subjects - by comparing the mathematical markers of the normal and abnormal brains.

In the schizophrenic brain, markers either remained the same as the controls (duplicate) or changed (nonduplicate) (Figure 4.5). By tallying the duplicate and nonduplicate markers, we can quantify the effect of schizophrenia on the living brain with one complexity (living brain) in play and on the post-mortem brain with two complexities in play (living brain + post-mortem artifacts).

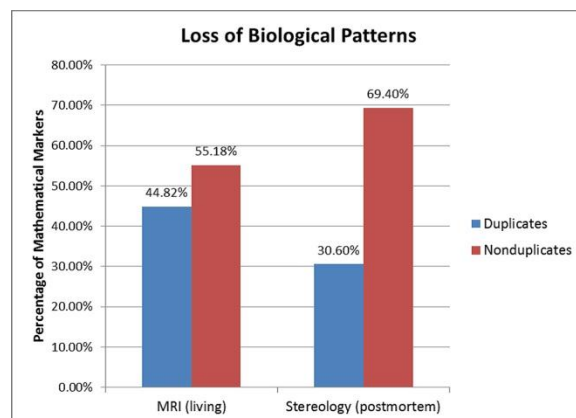


Figure 4.5 Schizophrenia changes quantitative patterns in the human brain by transforming duplicate markers into nonduplicates. Duplicates occur when the control and experimental markers are the same, indicating no change. In the living brain, schizophrenia decreased the percentage of duplicates to 44.82%, whereas in post-mortem brains the value fell to 30.60% (From Bolender, 2013). In the absence of post-mortem artifacts, both data sets – MRI and stereology -would be the same.

Figure 4.5 indicates that schizophrenia disrupts 55% of the markers in living brains, but this value climbs to 69% post-mortem. This takes us to a key point. The mathematical markers detecting changes in patterns post-mortem are likely to carry one set of distortions for normal brains and a different set for abnormal brains. This variability in the magnitude and direction of the post-mortem complexity introduces mathematical inconsistencies between the data sets of living and post-mortem brains, as seen in Figure 4.5. Were this not the case and both normal and abnormal brain suffered the same volume distortion, then both histograms would be the same. Obviously, they are not.

Next, we will consider the source of the artificial complexity (post-mortem artifacts).

4-4 Artificial Complexity

The artificial complexity includes all the post-mortem changes attributed to death and experimental methods. Although widely recognized, the conventional wisdom – stated or implied – prefers the view that such distortions play a minor or insignificant role and can be ignored. Consequently, most publications – including those that include stereological methods - do not include corrections for the artificial complexity. Since all the evidence presented thus far runs contrary to this conventional wisdom, we will continue the testing by comparing the markers of living and post-mortem brains.

Schizophrenia will serve as our next test case. The combined data pairs of the stereology and MRI databases gave 46,246 mathematical markers, which included control and experimental (schizophrenia) values for both volumes and numbers. Of this total, 33,130 markers were duplicates - 5,814 came from stereology and 27,316 from MRI. Of this group 2,763 duplicates were shared by the two data sources (stereology = MRI) with 709 coming from stereology and 2,054 from MRI. For volumes (stereology = MRI), the controls accounted for 156 duplicates and the experimentals 107. For details, see Bolender 2013.

Figure 4.6 summarizes these results. The compatibility between the stereology and MRI markers was only 5.6% for controls and 3.9% for experimentals. If, instead, we divide by 33,130 instead of 2,763, we get 0.5% for controls and 0.3% for experimentals. Once again, the mathematical markers show that living and post-mortem brains represent two very different structures quantitatively.

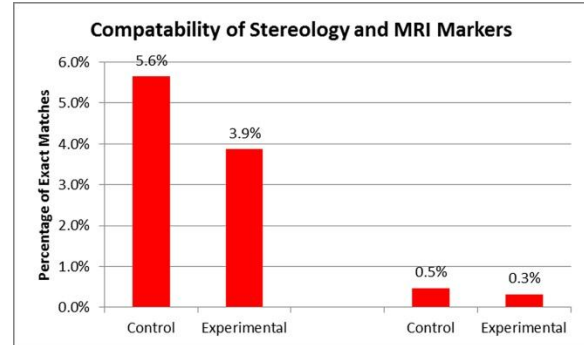


Figure 4.6 Just a small percentage of the post-mortem markers of stereology duplicate those in the living brain. Normal patients (control) are compared to those with schizophrenia (experimental). Since living and post-mortem brains define distinctly different phenotypes, they represent largely incompatible data sets (From Bolender, 2013).

If we view the compatible data set (5.6% and 3.9%) of Figure 4.6 with connectivity plots, the nature of the volume disruptions becomes apparent. Notice how the connectivity of the controls (Figure 4.7) largely disappears in brains with schizophrenia (Figures 4.8).

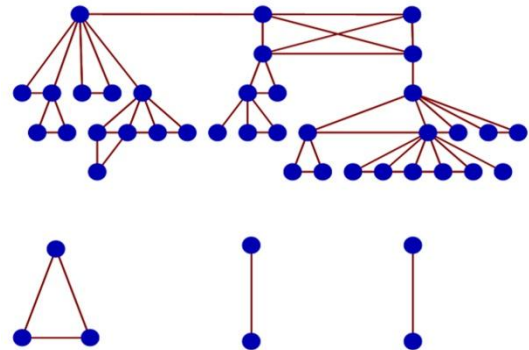


Figure 4.7 In the normal brain, these parts (represented by blue dots) form duplicate (identical) markers in both living and post-mortem brains. Evidence for a partial loss of connectivity appears as three isolated groups in the lower portion of the figure (From Bolender, 2013).

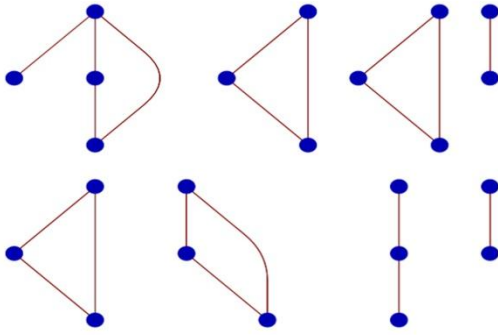


Figure 4.8 In brains diagnosed for schizophrenia, these parts (represented by blue dots) form duplicate markers in both living and post-mortem brains. Notice that the pattern of connectivity seen in the normal (Figure 4.6) has degraded to eight disconnected groups - most of which include subsets of a given part. This figure helps to explain why the post-mortem data of stereology lack the critical information needed to diagnose disorders of the brain (Move 13). The distortions in volume have largely overshadowed the original biological complexity (From Bolender, 2013).

Move 17: Can we determine the extent to which the parts of post-mortem brains differ from those of living brains?

Yes, mathematical markers can detect a range of quantitative differences that exist between living and post-mortem brains. Such comparisons indicate that the two data sets are largely incompatible.

4-5 Reality Check

Since living brains tell us one thing and post-mortem something else, we face a worst-case scenario. In post-mortem brains, the volume of each part may have increased (swollen), decreased (shrunken), or remained the same. Besides, the same part may distort differently in normal and abnormal brains. Given such a scenario, the standard stereological method of calculating absolute values with hierarchy equations - which typically ignores such distortions, may be introducing important errors – perhaps more often than not.

Whenever stereological data collected from post-mortem brains carry two distinct complexities, one attributed to biology and the other to experimental

methods, our data and interpretations can become ambiguous, confusing, and unreliable. Since we have been operating within this altered reality for some time, our options are limited. We can either restrict our data to volume independent estimates (fractionator-based cell counts), or figure out how to fix the problem.

Consider, for example, the situation that exists today in the neurosciences community. Data can come from two irreconcilable data sources, both of which share the same origin – the living brain. However, we routinely treat both data sets interchangeably, as if they were equal. This puts our goal of understanding the brain directly in conflict with our current policy for interpreting data. A best practices approach to this problem would be to reconcile the inconsistency by bringing both data sets into agreement. Otherwise, we face the unwelcome task of trying to defend a clearly indefensible position. To this end, move 17 discusses several approaches to this dilemma.

Move 18: If mathematical markers carry volume distortions, can we identify them and apply corrections?

Finding solutions to the problems created by volume distortions becomes an ongoing exercise, one that will require access to large amounts of published data and guidelines coming from the stereology community. The correction methods described below rely on the best data currently available.

4-6 Corrections for Post-mortem Data

The purpose of the following methods is to return post-mortem values to those that exist in the original, living material. In effect, we default to the living organism as the gold standard for interpreting research data. Defining a single gold standard is in keeping with complexity theory and its goal of approaching biology as a mathematical entity.

Method 1: The simplest solution to the problem is to make the before (living) and after (post-mortem) volumes the same, or nearly so. This requires

knowing the volumes of brain parts – in the same individual - before and after death. Such data generates correction factors – part by part - for the volume distortions:

$$\text{Correction Factor} = V(\text{part,before})/V(\text{part,after}) \quad (1)$$

$$\text{Correction Factor} = 80 \text{ mm}^3 / 70 \text{ mm}^3 = 1.143.$$

To fix a volume distortion in the post-mortem brain, simply multiply the post-mortem volume (after) by its correction factor.

$$V(\text{part,before}) = V(\text{part,after}) \times \text{Correction Factor} \quad (2)$$

$$V(\text{part,before}) = 70 \text{ mm}^3 \times 1.143 = 80 \text{ mm}^3.$$

Method 2: Since before and after estimates for the volumes of Method 1 are not yet available in the IBVD, we can estimate these correction factors using the currently available data set.

After identifying the parts that exist in both the stereology and MRI data tables, average values were calculated and used to evaluate equation (1). Figure 4.9 includes correction factors for 37 disrupted parts found in the control brains and Figure 4.10 does the same for brains with schizophrenia. These figures begin to explain why the mathematical markers of post-mortem and living brains shared so few duplicates (Figures 4.1, 4.2, 4.4, 4.5, 4.6). Moreover, they provide new information about the source, magnitude, and direction of the artificial complexity.

Notice that the histograms of Figures 4.9 and 4.10 identify correction factors that are both variable and subject to change. In human brains, the volume of each part typically responds uniquely to its post-mortem environment. This uniqueness, which represents a specific combination of post-mortem events, effectively scrambles the original biological data.

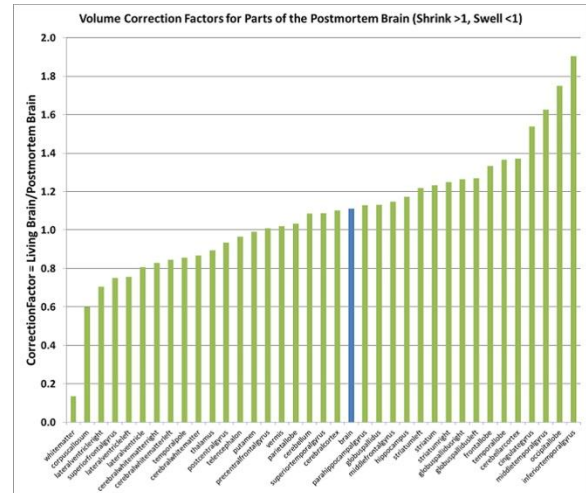


Figure 4.9 Volume correction factors for specific parts of the post-mortem human brain display a wide range of values. A value of 1 indicates no change, >1 shrinkage, and <1 swelling. The blue column identifies the brain, which requires a correction factor of 1.11 to account for shrinkage of about 11% (From Bolender, 2013).

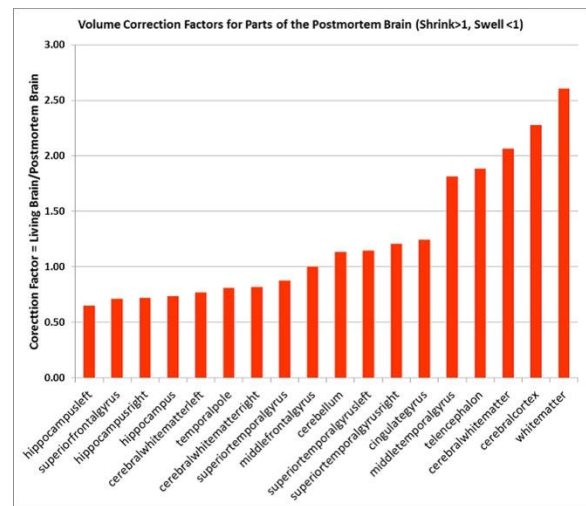


Figure 4.10 Volume corrections for parts of the post-mortem brain with schizophrenia display a wide range of values (From Bolender, 2013).

Method 3: Recall that our first attempts to diagnose schizophrenia failed repeatedly (Figures 4.1 and 4.2) because the MRI data standards were trying to diagnose the disorder in post-mortem brains that carried one complexity related to biology and an artificial complexity related to death and specimen preparation. If, however, we repeat the diagnosis after applying the correction factors for schizophrenia-

nia to remove the artificial complexity (Figure 4.10), we arrive at the correct diagnosis (Figure 4.11).

Now stereological data from post-mortem brains can work hand-in-hand with MRI data when playing the complexity game. This represents an important step because it means that both stereological data of post-mortem brains and MRI data of living brains can contribute to building the same parallel complexities. Recall that we will need stereology operating at the microscopic level to quantify the very small biological parts we will encounter as we make our way to the genome.

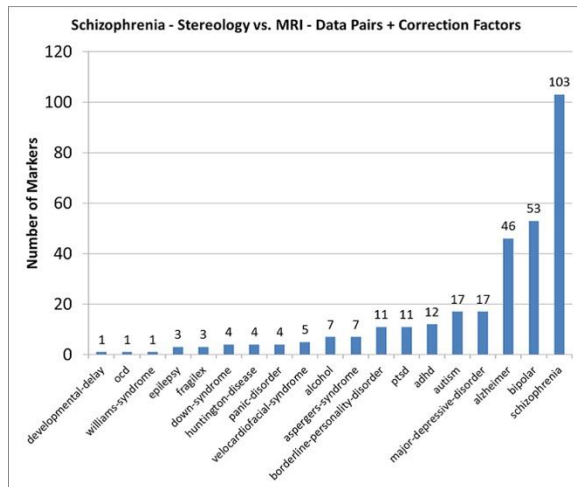


Figure 4.11 After applying the correction factors to the distorted volumes, we get the correct diagnosis - schizophrenia. When compared to the distribution of living brains (Figure 4.3), this post-mortem distribution shows striking similarities. Indeed, this distribution suggests that we can remove enough of the artificial complexity from post-mortem data to gain access to biology with the parallel complexity that remains. In effect, this puts the post-mortem data of the brain back in the game (From Bolender, 2013).

Method 4: The first three methods dealt exclusively with the human brain. To be inclusive, however, we prefer a solution to the distorted volumes problem – the artificial complexity - that requires only post-mortem data and generalizes to all types of parts, settings, and species.

If we estimate – using the unbiased sampling methods of stereology – the same parameter post-mortem with and without a volume distortion, then the ratio of the two should give us a correction factor for the distortion. Figure 4.12 plots cell counts –

estimated with the disector method (Sterio, 1984) - against the volumes of the parts containing the cells. R^2 's close to one can occur because both estimates share the same reference compartments and consequently the same volume distortions. Recall that the disector method estimates a numerical density (N/V), which when multiplied by a volume gives an absolute value (N). Such an estimate, which is volume dependent (vd), carries a volume distortion.

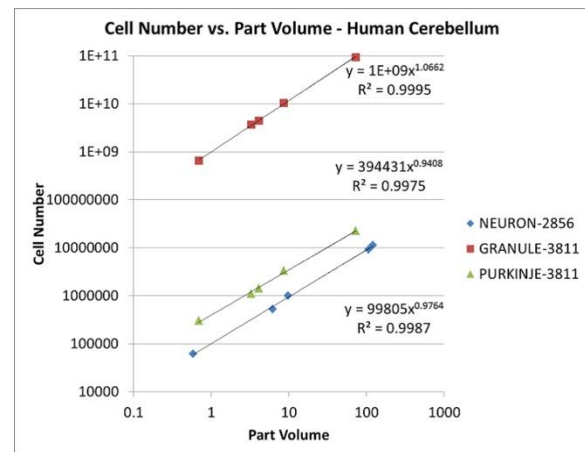


Figure 4.12 Cell numbers plotted against part volumes in the human cerebellum can display R^2 's close to 1.0 because both estimates share similar volume distortions. Adapted from Baker et al., 1999, Andersen and Pakkenberg, 2003 (From Bolender, 2013).

Recall that the fractionator method (Gundersen et al., 1988) gives a volume independent (vi) estimate for the total number of cells in a given biological part i - ($N(\text{cell}, vi)$). It becomes our lifeline because we can be reasonably confident that estimates for the total number of cells pre and post-mortem remain the same. Recall that cell counts are based on estimates of nuclear number.

The correction method requires several estimates. Using the same set of sections collected for fractionator-counting, the numerical density of the cells is estimated with the disector (Sterio, 1984) and the volume of the biological part containing the cells with the optical volume fractionator (Bolender and Charleston, 1993; Bolender et al., 1993). Since these estimates for cell number and part volume both

carry the same volume distortion, they supply the volume dependent (vd) estimate for cell number:

$$N(\text{cell}, \text{vd}) = V(\text{part}, \text{vd}) \times ((N_{\text{cell}}, \text{vd}) / V(\text{part}, \text{vd})). \quad (6)$$

Finally, to calculate a correction factor (CF) for a post-mortem part, we divide the volume independent cell count by the volume dependent one:

$$\text{CF} = N(\text{cell}, \text{vi}) / N(\text{cell}, \text{vd}). \quad (7)$$

To correct the volume of a post-mortem part, simply multiply the volume dependent part by the correction factor:

$$V(\text{part}, \text{vi}) = V(\text{part}, \text{vd}) \times \text{CF}. \quad (8)$$

This correction factor (CF) applies to all volume dependent estimates, including volume, surface, length, and number. Moreover, the method generalizes hierarchically up to and including the nucleus of the cell currently counted. Since each biological part carries a unique distortion (recall Figures 4.9 and 4.10), it has its own correction factor – estimated with equations 6 and 7. Unless shown to the contrary, the correction factor should be able to provide absolute data estimates largely free from the volume distortions associated with post-mortem material.

Move 18: If the mathematical markers carry volume distortions, can we identify them and apply corrections?

Yes, by combining the mathematical markers of two playing fields, we can work out corrections for the distortions of post-mortem data. Moreover, we can generalize these solutions across most parts of all organisms.

4-7 Summary of Chapter 4

Game 4 allowed us to reconcile two very different worlds of published data by applying complexity theory to a topic of fundamental importance - the quantitative relationship of data collected from the same part in living and nonliving states. In the absence of corrections for volume distortions, we discovered that a meaningful relationship between the two states could not exist.

The first principle to come from this game is that complexity theory requires living systems to serve as gold standards. In view of the results presented in this chapter, using biology-based standards to study biology becomes our most defensible position.

The next chapter pulls together the lessons we have learned thus far to figure out how to diagnose disorders of the brain. This will require new sets of parallel complexities along with a better understanding of how to manage large data sets and control experimental errors.

Chapter 5

Game 5 - Diagnosing Disorders of the Brain

Since we now have a better understanding of the relationship of pre to post-mortem data within the framework of two theory structures (reductionism and complexity), we can begin to address one of the fundamental challenges of clinical medicine – a data-driven approach to diagnosis. With this challenge comes the added promise of prediction, which, in a quantitative setting, becomes a connected set of diagnoses over time. The diagnosis of one patient becomes the prediction of another because in living systems local complexities retain their ability to generalize globally. By detecting these generalizations with mathematical markers, diagnosis and prediction represent two interpretations of the same complexity – differing only as a function of time. Such is the advantage of an objective approach to biology.

Diagnosis is at the heart of identifying and solving a disease or disorder. Once we know what is broken, we can look for a fix. By following mathematical patterns and pathways that exist in universal data sets, we can hunt for the culprits and evaluate our remedies.

5-1 Technology Shift

We know two things at the outset. In looking for a solution to the diagnosis problem, we can increase the specificity of a marker by increasing its number of variables and then figure out what combination of markers works by running a battery of exploratory tests. In both cases, however, we will have to make a technological shift from small data to big. Methods previously done manually will now require automation to operate on markers numbering in excess of 15,000,000. Finding a PC based solution to our big data problem turns out to be an interesting problem on its own. For our purposes here we will have to bring together the individual strengths of four software packages (PowerBuilder, Excel, Ac-

cess, and Mathematica) operating on 32 and 64-bit platforms.

A key finding of this chapter is a strategy for problem solving based on applying database filters to steer the results of a problem toward a best outcome. A solution to the diagnosis problem, for example, requires identifying a set of filtering algorithm that removes false positives and negatives from a parallel complexity. Detailed explanations of the procedures and worked examples are given elsewhere (Bolender, 2014).

5-2 Filtering Mathematical Markers

By defining a mathematical marker as a unit of biological complexity, it becomes a universal data type, which, in turn, we can use to quantify phenotypes. A database of such markers contains a large amount of complex information containing solutions to a wide range of problems. In such a database, however, these solutions intermingle with one another, leaving us with the task of extracting just the solution we want. Our goal, therefore, becomes one of applying filters to a database to produce the best parallel complexity for the current job.

Quadruplet Markers: A quadruplet marker is an alphanumeric string consisting of four named parts (A, B, C, D), each with a numerical values (X, Y, Z, Q). It defines the relationship of one part to another as a mathematical ratio (AX:BY:CZ:DQ). By increasing the number of variables in a marker, it increases its information content and specificity.

Switching to quadruplets, however, involves moving our data platform from small data to big. As shown in Figure 5.1, quadruplet markers quickly exceed the limits imposed by 32-bit Excel spreadsheets (2 GB of memory and $\leq 1,048,576$ rows of data). Since working with such mathematical markers includes shuttling large amounts of data back and forth between spreadsheets and databases, we will have to upgrade our software.

Quadruplet markers, however, introduce a host of new problems. Recall that diagnosis – based on mathematical markers - depends on matching unknown markers to known standards and tallying the results (Chapter 4 and Bolender, 2011-2013). This is done by adding an unknown set of markers to a table of known markers in a diagnosis database, sorting the markers alphabetically, and marking each duplicate marker (unknown=known standard) as it appears. Alternatively, we can transfer the database table to an Excel spreadsheet and use the conditional formatting tool to identify duplicates automatically. In practice, however, sorting a large data set automatically with a spreadsheet can take many hours.

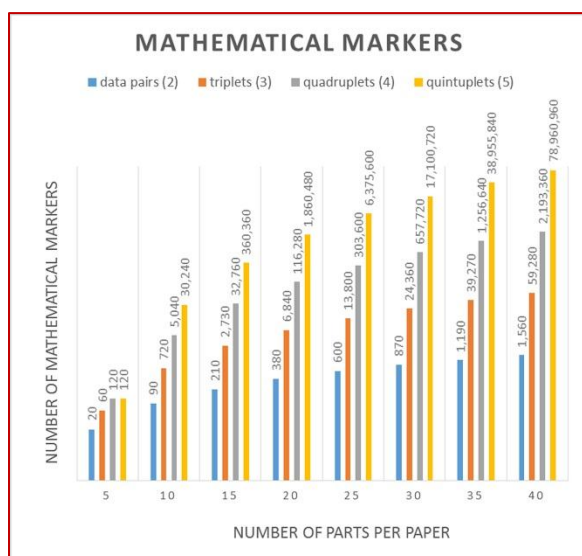


Figure 5.1 Original data sets increase the amount of information they contain by forming permutations. When expressed as mathematical markers these permutations define phenotypes both qualitatively (alpha string) and quantitatively (numeric string). Notice that the amount of data coming from a single paper can explode. For example, the same 20 parts can produce 380 data pairs, 6,840 triplets, 116,280 quadruplets, and 1,860,480 quintuplets. Each data set represents the phenotype as a set of patterns with a different degree of specificity (From Bolender, 2014).

The problems surrounding quadruplet markers can be resolved by shifting to a 64-bit platform, running Excel and Access together as a team, and using calculation templates (Bolender, 2014).

Making Known Markers: Testing the effectiveness of a diagnostic test requires two independent data sets – one for knowns and the other for unknowns. We begin with the volume data of a given paper in the IBVD and use the names of the parts to generate a list of quadruplets with the permutation function of Mathematica. Next, we import this list into an Excel worksheet as a text file (tab delimited) and use a template worksheet to associate each part (name) with its numerical value (volume). Once the ratios are calculated, they are assigned their decimal repertoire values. A template worksheet performs all the concatenations and calculations automatically, thereby producing a table of quadruplet markers. A diagnosis database is assembled by applying this procedure separately to control and experimental data sets – paper by paper.

Making Unknown Markers: To test the effectiveness of a diagnosis database, unknowns are prepared in the same way – one paper at a time - using the template described above. Since these data come from patients carrying both normal and abnormal markers, a false positive occurs whenever a normal marker in the unknown corresponds to an abnormal marker in the diagnosis database (control (unknown) = experimental (known)). To identify these false positives, we can run the unknown markers of a disorder against a database of normal markers and then delete the duplicates – the false positives.

Diagnosing disorders of the brain with a database of markers involves dealing with many more problems than one might first imagine (Bolender, 2012). In such a database, the same marker can occur one or more times, it can be unique or shared, and it can be a false positive. Developing a database of diagnostic markers requires an in-depth understanding of these issues. Trial runs and extensive testing become essential.

Since the earlier tests indicated that disorders were diagnosed correctly with shared markers (Bolender, 2012), we will make our first move with a database of these markers.

Move 19: Can shared markers based on quadruplet markers diagnose disorders of the brain correctly?

5-3 Test 1: Quadruplets (Shared Markers)

Starting with the original database of quadruplet markers derived from the IBVD, we apply filters to set the properties of the first diagnosis database. The filters remove duplicate markers from the same paper (control marker = experimental marker) and from the database (experimental marker = experimental marker) so that a given marker appears only once for a given disorder (Figure 5.2). The algorithm defines the diagnosis database for quadruplet markers, which is stored as a text file (Test1.txt).

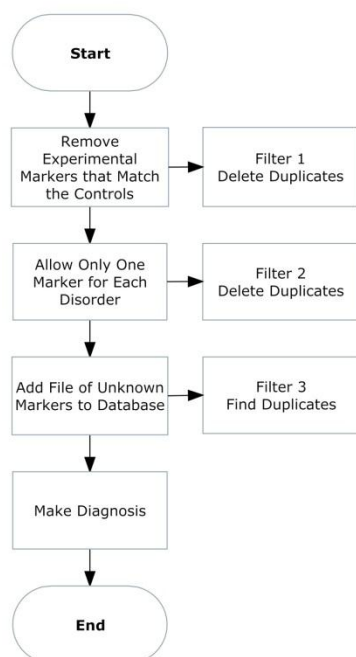


Figure 5.2 Algorithm 1 - Test 1 (From Bolender, 2014).

Diagnosing Unknowns (Shared Markers): To test the effectiveness of the algorithm, we import the database text file (test1.txt) into an Access database (64-bit), append a text file containing the markers of a test paper (unknown), look for matches (unknown = known), and tally the results. The diagnosis goes

to the disorder with the largest number of identified unknowns. Note that the Access database includes a routine for finding duplicates.

The problem with algorithm 1 is that it leaves a residue of false positives. Since the same marker can appear in several different disorders, the markers of one or more disorders can overwhelm those of another. In turn, this creates false positives that can lead to an incorrect diagnosis. The primary reason for running test 1 was to see if the quadruplet markers with their increased specificity could counteract the negative effect of the false positives.

Test Results: To test the effectiveness of this first database as a diagnostic tool, data from thirteen additional IBVD papers were translated into quadruplet markers and run – one by one as unknowns – against the knowns of the diagnosis database (Table 5.1). Note that the numbers identifying the papers correspond to those of the IBVD.

Table 5.1 The table includes the results of running the data of 13 unknowns (publications) – one at a time - against a collection of known standards, all of which came from the IBVD. A result can be correct (YES), incorrect (NO), tied (TIE), or nonexistent (no variables in play). In spite of the fact that more than 500,000 known markers were in play, the diagnosis was correct only about 50% of the time. Clearly, the correct diagnosis was repeatedly being overwhelmed by the data of other disorders (308, 329, 472, 587, 621, and 623). In two cases (308 and 472), the disorder being diagnosed failed to have even a single marker in play. Note that 5 of the 7 correct diagnoses came from the two disorders with the largest number of markers - schizophrenia (4) and bipolar (1) (From Bolender, 2014).

DIAGNOSIS OF UNKNOWN QUADRUPLT MARKERS - DIAGNOSIS DATABASE (DDB-2A-1) - TEST 1																
PAPER ID (IBVD)	UNKNOWN →	126	154	329	308	472	555	587	591	621	623	635	639	657		
NUMBER OF PARTS IN PLAY	→	12	10	9	7	6	22	7	7	18	7	13	9			
DISORDER	DUPLICATE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)														
ADHD	2,538	0	0	0	24	0	0	0	6	0	0	0	0	14	0	
AFFECTIVE-PSYCHOSIS	810	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ALCOHOL	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ALZHEIMER	9,815	0	0	6	6	0	23	69	6	234	30	70	0			
ASPERGERS-SYNDROME	1,566	54	0	0	0	0	20	7	0	0	0	0	14	0		
AUTISM	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
BIPOLAR	216,878	0	6	18	24	0	2	102	12	54	150	18	17	265		
BIPOLAR-ADHD	624	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
BORDERLINE-PERSONALITY-DISORDE	3,471	0	12	12	0	12	14	0	6	12	0	72	6	0		
DOWN-SYNDROME	239	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
EPILEPSY	107	0	0	24	0	30	0	0	66	24	0	0	0	18		
FRAGILEX	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
HUNTINGTON-DISEASE	102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
MAJOR-DEPRESSIVE-DISORDER	851	0	0	0	0	0	0	8	6	0	0	36	0	12		
OCD	102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
PANIC-DISORDER	462	0	0	0	0	0	0	129	0	2	150	0	0	0		
PRETERM	2,514	0	0	0	20	0	18	0	12	21	30	40	41	0		
PTSD	1,134	0	0	0	2	0	8	78	0	5	24	5	0	0		
SCHIZOPHRENIA	245,621	108	12	0	0	0	62	66	0	18	54	0	87	204		
SCHIZOTYPICAL-DISORDER	101,322	54	0	0	0	0	1	0	0	0	0	0	0	0		
VELOCARDIOFACIAL	1,416	0	0	52	0	96	0	0	36	78	0	0	0	24		
TOTAL MARKERS	589,944															
DIAGNOSIS	54% CORRECT	YES	YES	NO	NO	NO	YES	NO	YES	NO	NO	YES	YES	YES		
DIAGNOSIS	45% CORRECT W/O TIES	TIE							TIE							

The results of test 1 were disappointing in that the diagnosis was correct only 54% of the time; the rest (labeled NO) were false positives (46%) (Table 5.1). The table shows that several disorders masked the correct diagnosis (velocardiofacial, panic disorder, Down syndrome, and Alzheimer), but not schizophrenia. Although the increased specificity of the quadruplet markers played a role (no masking by schizophrenia), the number of parts not in play seemed to be a major limiting factor.

General Observations: The MRI papers from the IBVD supplied roughly 12,000,000 quadruplet markers for the control and experimental data sets. Eliminating duplicate markers (control = experimental) at the level of individual papers reduced this number to 4,796,416, and finally to 589,945 after deleting duplicates (experimental = experimental for a given disorder). This generated a diagnosis database for the known quadruplet markers. Notice that the filters of Figure 5.2 assure that a given mathematical marker occurs only once for a given disorder, but that the same maker can occur in different disorders.

Since a quadruplet marker contains four parts (names) with four connections (ratios), the fact that they existed across such a wide range of disorders seems remarkable in itself. Figure 5.3, for example, shows that the quadruplet database contains 2,538 markers for ADHD (red), but that ADHD shares its markers with at least 12 other disorders (blue). See Appendix I (Bolender, 2014) for histograms characterizing 21 different disorders of the brain. These histograms provide insight into the way biology manages and conserves its complexity.

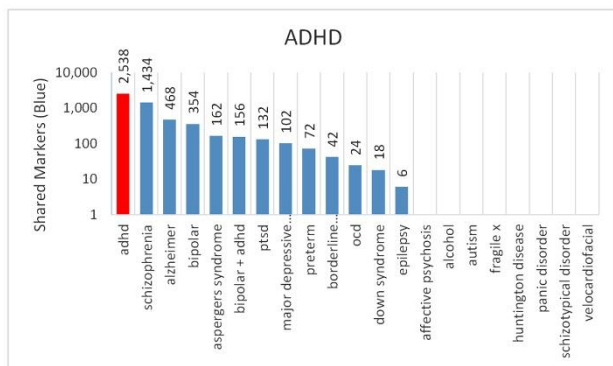


Figure 5.3 The diagnostic database contains 2,538 quadruplet markers for ADHD of which 1,434 also occur in schizophrenia, 468 in Alzheimer, etc. (From Bolender, 2014).

Figure 5.4 itemizes the frequency distribution of the quadruplet markers – by disorder - for the database used in Test 1. Notice that the markers range in number from 24 to 245,621, that schizophrenia and bipolar disorder account for 96% of the total, and that the Y-axis is logarithmic.

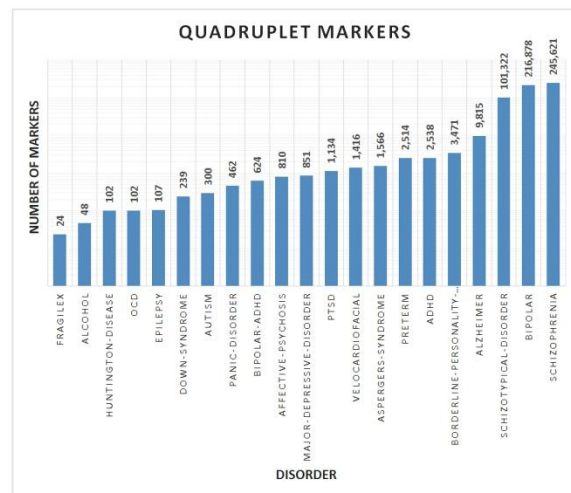


Figure 5.4 The histogram illustrates the frequency distribution of quadruplet markers in the diagnosis database across 21 disorders of the brain. Notice that most of the markers belong to schizophrenia and bipolar disorder (From Bolender, 2014).

Apparently, the test failed because it attempted to compare known and unknown samples that were poorly matched. If true, then increasing the amount and mix of data in the diagnosis database (knowns) should improve the result. Test 2 considers this possibility by shifting from quadruplet to triplet markers.

Move 19: Can shared markers based on quadruplet markers diagnose disorders of the brain correctly?

No, shared markers result in only about a 50% success rate because the markers of one disorder can mask those of another. Such a result illustrates the power of false positives.

By switching from quadruplet markers to triplets, the next move will determine if increasing the number of shared markers in play improves the diagnosis.

Move 20: Can shared markers based on triplet markers diagnose disorders of the brain correctly?

5-4 Test 2: Triplets (Shared Markers)

This test consisted of downsizing the quadruplet markers of test 1 to triplets (AX:BY:CZ), while at the same time increasing the number of IBVD papers contributing markers to both the known and unknown data sets (Figure 5.5). Notice in Table 5.2, however, that test 2 also failed at about the same level as test 1. The diagnosis was correct only 57% of the time. In addition, the triplet markers displayed a substantial loss of specificity, as shown by the strong masking effect by schizophrenia. Remove the masking effect, however, and the success of the test jumps to 86%.

Table 5.2 The diagnosis database of Test 2 included the same collection of parts used in Test 1, but this time they were used to produce triplet markers. The table, which shows a strong masking effect by schizophrenia, produced a diagnostic score of only 57%. When the masking effects of schizophrenia data were removed from the analysis, the score improved to 86% (From Bolender, 2014).

DIAGNOSIS OF UNKNOWN TRIPLET MARKERS - (T-DOB-3A) - TEST 2																										
PAPER ID (IBVD)	UNKNOWN(S) →	50	126	132	329	358	482	485	536	555	587	590	591	621	626	635	639A	639	657	667	724	777				
NUMBER OF PARTS IN PLAY	→	27	12	9	9	8	7	9	11	22	7	6	7	7	10	7	13	13	9	12	7					
DISORDER	MARKERS IN DB ↓	UNKNOWN MARKERS IDENTIFIED BY DISORDER (GREEN=DIAGNOSIS)																								
ADHD	249	29	0	0	0	0	2	0	0	0	0	0	0	2	0	0	0	7	2	2	0	0	0	0	0	
AFFECTIVE-PSYCHOSIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ALCOHOL	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ALZHEIMER	4,391	26	0	0	0	0	0	0	0	46	7	1	1	0	2	0	15	12	1	0	0	0	0	0	0	
ASPERGERS SYNDROME	221	12	0	0	0	0	2	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
AUTISM	113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
BIPOLAR	17,529	23	1	0	10	1	21	0	0	8	15	1	4	10	7	1	0	4	27	0	0	0	0	0	0	
BIPOLAR-ADHD	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
BORDERLINE PERSONALITY DISORDER	271	12	1	0	1	1	1	0	0	8	1	0	1	1	3	1	6	0	1	0	0	0	0	0	0	
DEVELOPMENTAL-DELAY	124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
DOWN SYNDROME	15	0	0	0	2	0	4	0	0	0	0	0	0	15	2	0	0	0	0	0	0	0	0	0	0	
EPILEPSY	18	0	0	0	7	0	6	0	0	0	0	0	12	8	2	0	0	0	7	0	0	0	0	0	0	
FRAGILEX	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
HUNTINGTON-DISEASE	236	0	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
MAJOR-DEPRESSIVE-DISORDER	3,381	11	0	0	0	0	2	0	8	0	14	0	0	0	0	0	6	3	0	0	0	0	0	0	0	
OCD	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
PANIC DISORDER	40	8	0	0	1	0	0	0	0	0	13	1	1	0	0	0	0	0	1	0	0	0	0	0	0	
PRETERM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
PTSD	103	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
SCHIZOPHRENIA	99,166	446	169	4	10	17	26	25	1	174	46	12	16	40	49	20	85	22	93	7	4	1				
SCHIZOTYPICAL DISORDER	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
VELLOCARDIOFACIAL	155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
WILLIAMS SYNDROME	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
TOTAL MARKERS	126,212																									
DIAGNOSIS	57% CORRECT	NO	YES	YES	NO	YES	NO	YES	YES	YES	NO	YES	YES	NO	YES	NO	NO	NO	YES	YES	YES	NO	YES			
DIAGNOSIS	56% CORRECT W/OT TIES	TIE																								
MASKING BY SCHIZOPHRENIA	→	X		X		X		X				X					X	X	X				X			
DIAGNOSIS W/O SCHIZOPHRENIA	86% CORRECT	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES			
DIAGNOSIS	86% CORRECT W/OT TIES	TIE																								

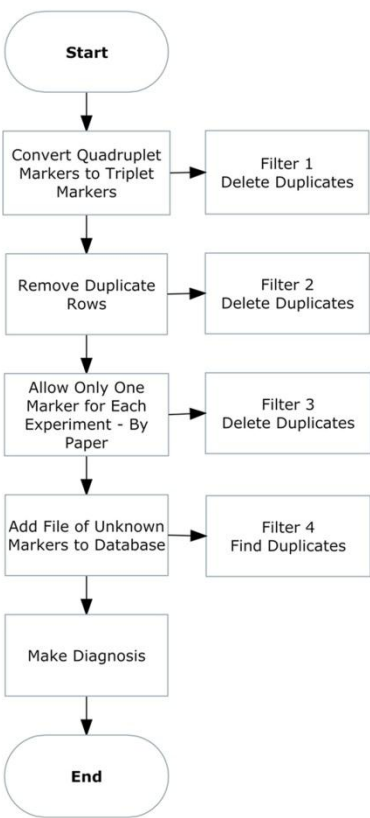


Figure 5.5 Algorithm 2 - Test 2 (From Bolender, 2014).

Move 20: Can shared markers based on triplet markers diagnose disorders of the brain correctly?

No, shared markers result in only about a 50% success rate because the markers of one disorder continue to mask those of others. The problem of false positives remains.

Taken together, the results of tests 1 and 2 suggest that a database containing just shared markers offers little promise as a diagnostic tool. Accordingly, test 3 will try unique markers.

Move 21: Can unique triplet markers diagnose schizophrenia correctly?

5-5 Test 3: Triplets (Unique Markers)

Next, we operate on data derived exclusively from patients with schizophrenia. Recall that schizophrenia carries two types of markers, those that it shares with other disorders and those unique to schizophrenia. When we select only the unique markers (the ones that occur only once) from the diagnosis database of triplets described in Test 2, we get 83,305 for schizophrenia (Figure 5.6). If, in turn, we run several unknowns against this new set of unique markers (knowns), the effectiveness of the diagnostic method jumps to 100% (Table 5.3).

Table 5.3 Run the unknown markers against the database of markers unique to schizophrenia and the unknowns are diagnosed correctly (From Bolender, 2014).

DIAGNOSIS OF UNKNOWN TRIPLET MARKERS - (T-DDB-3B) - TEST 3 MARKERS UNIQUE TO SCHIZOPHRENIA														
PAPER ID (IBVD)	UNKNOWN →	126	154	358	555	587	621	623	639	657	667	669	777	
NUMBER OF PARTS IN PLAY	→	12	10	8	22	7	7	18	13	9	12	9	7	
DISORDER	MARKERS IN DB ↓	UNKNOWN MARKERS IDENTIFIED AS SCHIZOPHRENIA												
SCHIZOPHRENIA MARKERS	83,305	141	375	14	144	34	8	39	27	64	7	5	1	
DIAGNOSIS	100% CORRECT	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	

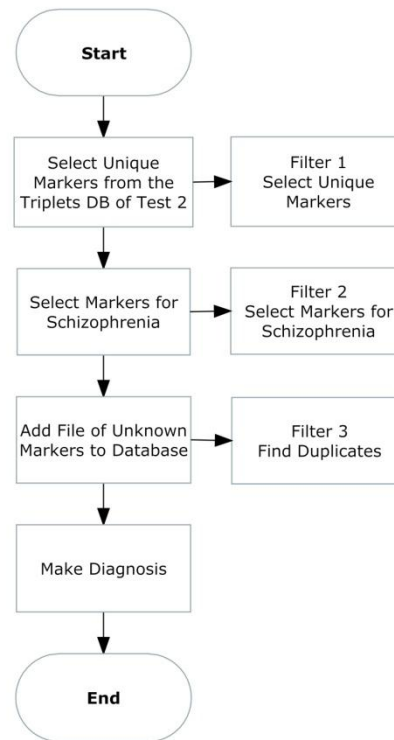


Figure 5.6 Algorithm - Test 3 (From Bolender, 2014).

Move 21: Can unique triplet markers diagnose schizophrenia correctly?

Yes, a database of triplet markers unique to schizophrenia can diagnose unknown markers correctly 100% of the time.

The problem with the test is that it succeeds only when the unknowns come from patients with schizophrenia. However, the results suggest that a solution to the diagnosis problem at least seems possible. The next test (Move 21) directs the question to all disorders and quadruplet markers.

Move 22: Can a database of unique quadruplet markers diagnose disorders correctly?

5-6 Test 4: Quadruplets (Unique Markers)

We return to the quadruplets database and select only the unique markers - those that appear only once in the database (Figure 5.7). When we run several unknowns against the unique knowns of this database, however, the promising results seen in Table 5.3 fail to appear. Table 5.4 shows that the database of unique markers has a success rate of only 8%. Moreover, the markers of four papers (308, 587, 621, and 657) were not even in play, as indicated by the (0). Although the markers are unique in the quadruplet database of knowns, they were not unique in the unknowns because the same marker occurs in more than one disorder. In effect, the unknowns are producing false positives.

Table 5.4 The database of unique quadruplet markers was not effective because it shares its unique markers with more than one of the unknown disorders. In effect, the known markers are unique to the knowns but not to the unknowns (From Bolender, 2014).

DIAGNOSIS OF UNKNOWN QUADRUPLLET MARKERS - DIAGNOSIS DATABASE (QUADS-UNIQUE) - TEST 4														
PAPER ID (IBVD)	UNKNOWN →	126	154	308	329	472	555	587	591	621	623	635	639	657
NUMBER OF PARTS IN PLAY	→	12	10	7	9	6	22	7	7	7	18	7	13	9
DISORDER	UNIQUE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)												
ADHD	240,304	42	0	30	0	0	22	0	6	0	0	0	132	0
AFFECTIVE-PSYCHOSIS	1,008	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	1,212	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	587,743	0	12	0	0	0	442	210	0	0	318	6	30	0
ASPERGERS-SYNDROME	337,914	42	0	0	0	0	22	0	0	0	0	0	0	0
AUTISM	10,797	0	0	0	0	0	12	0	0	0	0	0	0	0
BIPOLAR	770,306	12	24	48	192	0	40	90	6	54	102	12	312	0
BIPOLAR-ADHD	450	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE- PERSONALITY-DISORDER	50,679	0	0	0	0	0	402	0	0	0	0	18	0	0
DOWN-SYNDROME	559	0	0	54	12	18	0	0	42	208	0	0	0	0
EPILEPSY	433	0	0	42	12	12	0	0	204	12	0	0	0	0
FRAGILEX	6,372	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	22,410	0	0	0	0	0	0	0	0	0	0	0	0	0
MAJOR-DEPRESSIVE-DISORDER	44,325	0	6	0	0	0	16	0	0	0	0	6	0	0
OCD	38,685	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	15,006	0	0	0	0	0	36	0	0	24	0	0	0	0
PRETERM	49,836	0	0	0	24	18	50	0	18	0	0	18	0	0
PTSD	64,362	12	0	0	0	0	826	0	0	0	24	6	0	0
SCHIZOPHRENIA	1,189,381	624	6	0	0	0	194	0	6	0	114	24	36	0
SCHIZOTYPICAL-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VELOCARDIOFACIAL	73,446	0	0	0	0	132	30	0	18	18	0	42	0	0
TOTAL MARKERS	3,505,228	732	48	174	240	180	2056	336	282	310	582	114	528	0
DIAGNOSIS	8% CORRECT	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

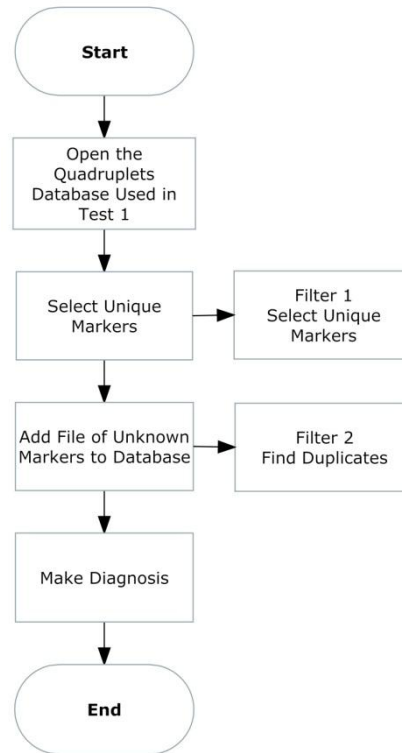


Figure 5.7 Algorithm - Test 4 (From Bolender, 2014).

Move 22: Can a database of unique quadruplet markers diagnose disorders correctly?

No, this database can diagnose unknown markers correctly only 8% of the time.

The results thus far suggest that we are in trouble because our attempts to solve the diagnosis problem fail repeatedly. Complexity theory, however, reassures us that every problem and solution intrinsic to biology can exist in a parallel complexity - provided we set up the problem correctly. We simply need to rethink our approach.

What do we know so far? We know that a diagnosis succeeds when the all the markers (known and unknown) are unique (Table 5.3), but fails when one or both of the markers – known and unknown - are shared (Tables 5.1, 5.2, and 5.4). Tests 1 and 2 failed because they used just shared markers. Test 3 succeeded because both the known and unknown markers were unique. Test 4 failed because one set

of markers was unique (knowns), but the other set was shared (unknowns).

If we study these results carefully, they provide all the clues required to solve the problem. We need to apply a set of filters that prevent or minimize sharing within - but not between - the known and unknown markers. In effect, we can diagnose a disorder of the brain by matching unknown to known markers, provided such markers are unique to the individual known and unknown data sets. In Test 5, we take the next step by applying a filtering algorithm that improves the number of successful outcomes.

Move 23: Can additional filters improve the success of a database of unique quadruplet markers in diagnosing disorders correctly?

5-7 Test 5: Quadruplets (Unique Markers)

Now, we can begin to zero in on a solution. Test 4 uses one unique filter, whereas Test 5 uses two (Figure 5.8). The first filter of Test 5 selects for unique markers, whereas the second filter selects for markers unique to a given disorder – paper by paper. The resulting markers serve as the knowns in the diagnosis database of Test 5. Table 5.5 indicates that this new filtering algorithm leads to a better outcome, given the resulting score of 80%. Notice that three of the unknowns (472, 587, and 657) were out of play (OOP) because Filter 3 found no matches. Furthermore, the unknown markers of papers 308 and 621 lead to an incorrect diagnosis of epilepsy and that the correct diagnosis was sometimes out of play, as indicated by the absence of duplicates (0). These results tell us that we are reducing, but not eliminating issues related to sampling, data compatibility, false positives, and false negatives. If we remove the three OOP unknowns (false negatives), then the diagnostic algorithm works successfully 100% of the time.

Table 5.5 By increasing the uniqueness of the markers, we increase their ability to diagnose disorders correctly. Removing papers that are out of play (OOP) improved the results

(From Bolender, 2014).

PAPER ID (IBVD)	UNKNOWN ¹ →	126	154	308	329	472	555	587	591	621	623	635	639	657
NUMBER OF PARTS IN PLAY	→	12	10	7	9	6	22	7	7	7	18	7	13	9
DISORDER	UNIQUE MARKERS ↓	UNKNOWN MARKERS IDENTIFIED (GREEN=DIAGNOSIS)												
ADHD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AFFECTIVE-PSYCHOSIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALCOHOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALZHEIMER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASPERGERS-SYNDROME	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUTISM	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BIPOLAR	312	0	0	0	0	0	0	0	0	0	0	0	312	0
BIPOLAR-ADHD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BORDERLINE-PERSONALITY-DISORDER	18	0	0	0	0	0	0	0	0	0	18	0	0	0
DOWN-SYNDROME	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EPILEPSY	204	0	0	36	12	0	0	204	12	0	0	0	0	0
FRAGILEX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HUNTINGTON-DISEASE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MAJOR-DEPRESSIVE-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OCD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PANIC-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRETERM	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PTSD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCHIZOPHRENIA	932	625	6	0	0	0	194	0	0	0	114	0	4	0
SCHIZOTYPICAL-DISORDER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VELOCARDIOFACIAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL MARKERS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DIAGNOSIS 1 (QUADS-UNIQUE ¹)	80% CORRECT	YES	YES	NO	YES	OOP	YES	OOP	YES	NO	YES	YES	YES	OOP
DIAGNOSIS 2 (QUADS-UNIQUE ²)	100% CORRECT	YES	YES	OOP	YES	OOP	YES	OOP	YES	OOP	YES	YES	YES	OOP

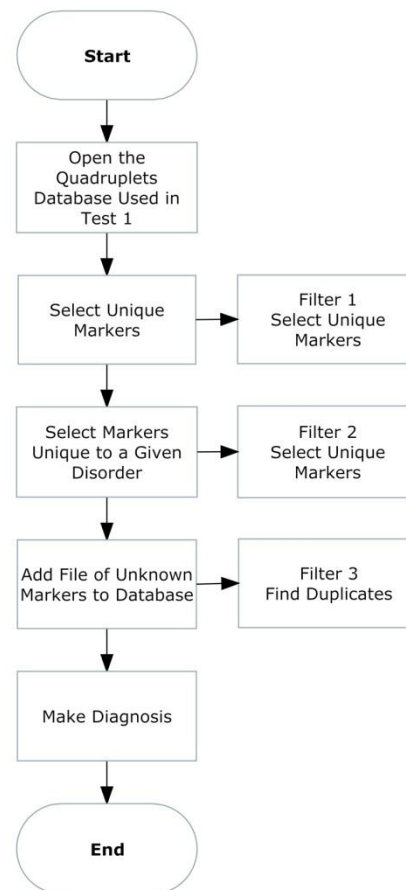


Figure 5.8 Algorithm Test 5 (From Bolender, 2014).

Move 23: Can additional filters improve the success of a database of unique quadruplet markers in diagnosing disorders correctly?

Yes, by adding an extra filter, this database can now diagnose unknown markers correctly 80% of the time.

The remaining tests show that a success rate of 100% becomes possible if we introduce additional filters and define the boundaries of the diagnostic space. In effect, a success rate of 100% depends on putting the markers in a data cage so that we can remove all the false positives and negatives. At first, it may appear that we are gaming the system, but in reality, we are building a foundation for the predictive modeling of the disease process. By getting all the false positives and negatives under control, we can then predict the likelihood of a correct diagnosis for patient data existing outside of the cage. If, for example, a diagnosis is incorrect for a given patient, such a mistake can occur only once because the patient's data will be added to the cage.

Move 24: Can unique quadruplet markers diagnose disorders correctly 100% of the time correctly?

5-8 Test 6: Quadruplets (Unique Markers)

We can arrive at a diagnostic score of 100% by dealing successfully with all the outstanding issues (Figure 5.6). By designing a database wherein all the known and unknown markers are unique and each marker can assume the role of both known and unknown, the disrupting factors disappear. Notice that this curious solution was driven entirely by the complexity itself. A diagnosis database capable of a 100% success rate requires a closed system, wherein only those markers coming from the IBVD are in play. However, markers derived from sources ex-

ternal to the IBVD can be expected to approach the 100% level as the properties of markers in the diagnosis database approaches that of the general population.

Since the results of the test indicated that only unique mathematical markers could give the correct results 100% of the time, the diagnosis database for quadruplets (MRI_Q_DIAG_100) contains just such markers (Figure 5.9). Table 5.6 summarizes the composition of this database, which contains data from 75 papers and 3.6 million unique markers. A marker generated from any one of these 75 papers and run against this database, will give a correct diagnosis.

Table 5.6 With the appropriate filters applied, a quadruplet database of unique markers diagnoses a disorder correctly 100% of the time (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%

Table 5.6 offers a gentle wake-up call. If the IBVD is representative of the clinical literature, then three or fewer papers are representing 86% (19/22) of the disorders. Such small sample sizes will at some point compromise our ability to diagnose and predict outcomes at the 100% level – but only when we stray beyond the boundaries of our closed system (data cage). When this occurs, a diagnosis reverts

to a prediction with a probability yet to be determined.

The inescapable conclusion to come from Table 5.6 is that interacting with biology by means of a parallel complexity is going to involve extremely large data sets especially when clinical diagnosis is the goal. We can expect such diagnostic and predictive databases to become fundamental to our health care systems.

Move 24: Can unique quadruplet markers diagnose disorders correctly 100% of the time correctly?

Yes, by removing false positives with filters, quadruplet markers can diagnose disorders correctly 100% of the time – in a closed system.

Move 25: Can unique triplet markers diagnose disorders correctly 100% of the time correctly?

5-9 Test 7: Triplets (Unique Markers)

This test applies the procedure described for the quadruplet markers of Test 6 to triplet markers (Figure 5.10). Once again, the diagnosis of unknown markers was correct 100% of the time (Table 5.7).

By adding one more filter, we can minimize the effect of false positives that may occur when the diagnosis database is used to predict a disorder with an unknown set of markers – existing outside of the data cage. This includes, for example, data that would come from a single patient. Recall that a marker of a disorder becomes a false positive whenever a control marker duplicates it. These duplications occur at two levels - papers and databases. We can remove false positives ($C=E$) from a given paper by identifying duplicates between normal (C) and abnormal (E) markers. Once the diagnosis database is built, it can be run against the original database of normal markers to delete the remaining false positives ($C=E$ for all papers) in the database. This database filter, for example, removed an additional 31,275 false positives from the MRI-T-Diag-100 database of Test 7. This additional step reminds us of the behavior of a complex system, wherein both local and global issues are always in play.

Notice in Tables 5.6 and 5.7 that the markers characterizing 22-27 disorders of the brain came from a relatively small number of papers - 75 for quadruplets and 117 for triplets. Given the software tools

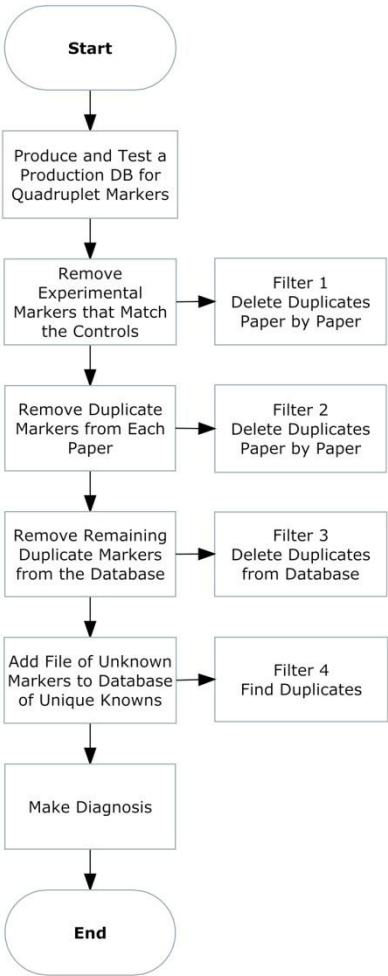


Figure 5.9 Algorithm - Test 6 (From Bolender, 2014).

included in the 2014 software package, the task of increasing substantially the number of papers in play becomes a reasonably straightforward task.

Table 5.7 When filtered appropriately, a triplet database of unique markers can diagnose a disorder correctly 100% of the time (From Bolender, 2014).

TRIPOLET 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
TRIPOLET TOTALS	277,039	117	100%

Move 25: Can unique triplet markers diagnose disorders correctly 100% of the time correctly?

Yes, by removing false positives with filters triplet markers can diagnose disorders correctly 100% of the time – in a closed system.

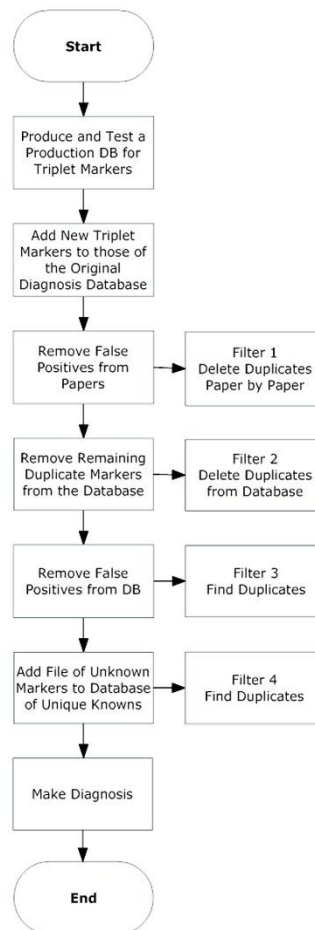


Figure 5.10 Algorithm - Test 7 (From Bolender, 2014).

5-10 Summary of Chapter 5

The solution to the diagnosis problem described herein may seem curious at first reading because it follows an unfamiliar set of rules. Instead of using the signs and symptoms of a disorder to make a diagnosis, it uses a set of unique markers taken from diagnosed patients to define a given disorder as a unique phenotype. By closing the system, we optimized both accuracy and precision – simultaneously. In effect, we used a parallel complexity to create a data cage wherein all the disorders in play will always be diagnosed correctly - 100% of the time.

Given biology with its nested complexities and redundant connectivity, a given variable (e.g., mathematical marker) can be subject to wide range of influences. This explains – at least in part - why the

diagnosis problem was so difficult to solve. False positives and negatives were interfering with our goal of 100% effectiveness because we did not know how to manage the instability attributed to the unknown variables. By putting our parallel complexity into a data cage, we eliminated this instability by identifying and removing all the false positives and negatives with filters. Since the data cages have no practical upper limits on the amount of data they can contain, the data sets can approach a global reality. In effect, this software device allows us to isolate and operate on a complex data set without suffering the limits imposed by outside disturbances. This newly acquired skill creates new opportunities. When, for example, we want to look at the same disorder in different subsets, populations, or stages of development, data cages will allow us to ferret out subtle differences. Furthermore, by generating millions of data cages for every type of phenotype imaginable, we can redefine diagnosis and prediction as a connected set. In effect, data cages will allow us to explore even the most complicated complexities.

Mathematical markers create an intriguing view of biology by demystifying the nature of its complexity. Although they begin as one-dimensional strings of parts and connections, we have seen that they readily concatenate to form two-dimensional surfaces, and three-dimensional networks. This suggests that biology is operating in n -dimensional

space. This becomes a useful construct because we can use it to associate a given task with a dimension. Diagnosis and prediction, for example, require one-dimensional strings, whereas identifying the background of a disease requires information existing in higher dimensions. In effect, we have a new strategy. Since everything is connected, mathematical markers serve to define a coherent infrastructure for complex problem solving.

In clinical diagnosis, the importance of connectivity begins to assume a much broader meaning. Changes in the brain can prompt changes to occur in the periphery and vice versa (Agostini et al., 2012; Borson et al., 2008; Cecil et al., 2008; Clarence et al., 1999; Guido et al., 2013; Herting et al., 2014; Khan et al., 2012; Nagai et al., 2010; Pérez-Dueñas et al., 2006; Strassburger et al., 1997; Tiehuis et al., 2008). This means that data collected at one location in an individual can be used to diagnose or predict an event at another. Accordingly, our ability to quantify human phenotypes triggers a host of new opportunities.

By combining the expertise of leading clinicians with MRI data to form parallel complexities, we now know how to recruit the biology literature as a reliable problem solver. In Game 6, we will attempt to extend the strategy developed for diagnosis to the disease process.

Chapter 6

Game 6 – The Disease Process

How does biology – in keeping with its rules of complexity – produce disease in the human brain? Whatever the cause, biology replaces a normal complement of parts and connections with an abnormal one (Chapters 3-5). By changing the phenotype of the brain, new properties and symptoms appear that we regard as abnormal. Our goal in this chapter will be to use parallel complexities to identify patterns that begin to explain the disease process at the level of a quantitative phenotype.

The game is simple. We will start with the big picture – created by combining the mathematical markers of twenty-one disorders – and then unfold it stepwise to discover what changes occur, what parts play major roles, how disorders overlap, and how symptoms relate to markers.

6-1 Unfolding the Complexity of Disease

We can study the disease process as it relates to disorders of the brain by assembling a parallel complexity from the disorders available to us in the IBVD. Since we will be looking for shared patterns, the database defining our parallel complexities will include only shared (duplicate) markers.

What will we learn from these shared markers? They can show us where the abnormal patterns appear, identify relationships of patterns to diseases, and allow us to observe an individual disease as a complexity embedded within the larger complexity defined by the disease process. By combining twenty-one different disorders from the IBVD into a composite brain, we begin with a global view of the disease process.

Starting with this composite brain as our parallel complexity, we will peel it apart objectively with mathematical markers to examine the design of each disorder and its relationship to the disease process. We will discover that a modular design is the com-

mon thread running through all the disorders, one that identifies yet another first principle. Biology, it would appear, creates a collection of building blocks and then deploys them in different ways to create normal and abnormal brains. Not surprising, biology is playing a game similar to the one nature plays with the periodic table of elements – good ideas tend to encourage emulation.

Since large data sets and patterns will be in play, most of the results are presented graphically. To this end, we will apply a new algorithm from Mathematica 10 (Wolfram Research) - the CommunityGraphPlot – to parse our composite brain.

Move 26: Do disorders of the brain adhere to a modular process?

6-2 Generalizing Disorders with Modular Markers

According to complexity theory, a generalization exists when the same pattern occurs both locally and globally - repeatedly. Since mathematical markers contain a well-defined set of parts and connections, we will use them as a proxy for the modular building blocks of biology. Both quadruplet and triplet markers will be in play.

Quadruplet Markers: Our first move verifies the existence of such modules defined either as quadruplet (Table 6.1) or triplet markers (Figure 6.1).

In Table 6.1, quadruplet markers display 10 different duplicate groups, ranging from 2 to 11 occurrences per group. Even though the alphanumeric string of the quadruplet markers contain 8 variables (AX:BY:CZ:DQ), 21% of the markers in the quadruplets database form duplicates globally. This represents more than two million markers.

Notice in Table 6.1 that the percentage columns to the right identify a conspicuous shift in the frequency distribution of the duplicate groups from 2 copies

to 3 and 4. This tells us that the disease process involves an increase in connectivity. Not surprisingly, this observation is exactly opposite to what we found earlier for post-mortem data in Figures 4.7 and 4.8. This inconsistency can be explained by the presence of multiple complexities in play (Figure 3.1).

Such inconsistencies serve to remind us of the risk involved in extrapolating post-mortem observations to living systems (Chapter 4). More importantly, however, we are beginning to understand why so many inconsistencies can exist in the biology literature and why experiments can be so difficult to reproduce. Unless we have enough of the right information, our experimental outcomes all too often end up carrying an unacceptably high degree of risk. Wrong information can quickly turn the results of an experiment upside down.

Table 6.1 The distributions of quadruplet markers suggest that the brain responds to the disease process by increasing 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 (From Bolender, 2014).

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%

Triplet Markers: The database of triplet markers included 381,476 duplicate markers, which represented 47.2% of the total population. The global distribution of these duplicate markers ranged from 2 to 64 per group (Figure 6.1).

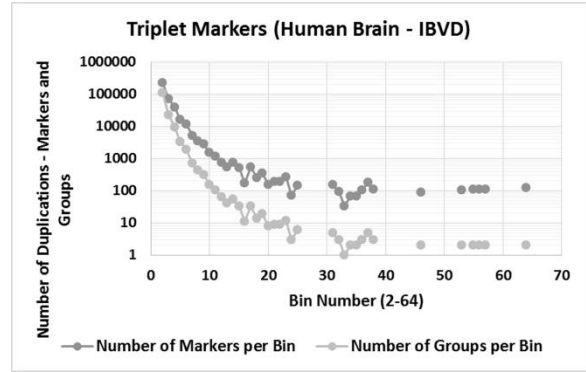


Figure 6.1 The distribution of triplet markers (C+E) shows duplications ranging from 2 per group to 64 (From Bolender, 2014).

Move 26: Do disorders of the brain adhere to a modular process?

Yes, human brains in health and disease display large number of duplicate modules locally and globally.

Since Table 6.1 and Figure 6.1 verify that modules (markers) exist in large quantities at the global level, we can approach the disease process as a mathematical puzzle. Starting with disorders of the brain as a general complexity, we can tease it apart to uncover specifics of the disease process.

Move 27: Can we unfold the complexity of brain disorders into well-defined communities of markers and disorders?

6-3 Finding Communities of Disorders

Using our database of duplicate markers, we can create a parallel complexity for an imaginary brain suffering simultaneously from 21 different disorders (Figure 6.2). In turn, we can unfold the nested complexity of this brain to discover how these disorders are related (Figures 6.23-6.8). The unfolding process continues until only clusters of two disorders remain.

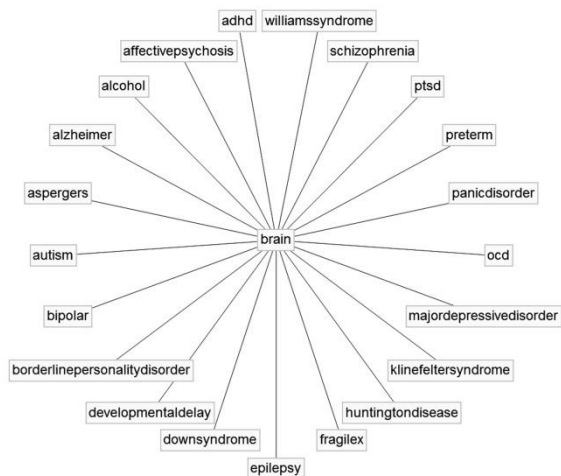


Figure 6.2 A database of duplicate mathematical markers becomes a parallel complexity representing twenty-one disorders of the human brain (From Bolender, 2015).

Step 1: When applied to the entire database of duplicate mathematical markers (Figure 6.2), the CommunityGraphPlot (Mathematica) identifies five distinct clusters (Figure 6.3), four of which contain closely related disorders. The patterns displayed by the dark blue lines (connectivity) and the dots (markers) suggest that the disorders are all connected and that they share many of the same abnormal markers.

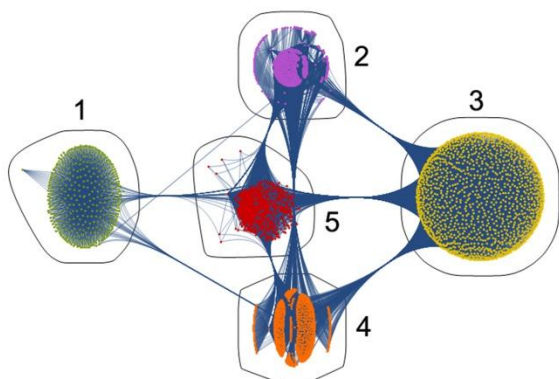


Figure 6.3 Step 1: The collection of shared mathematical markers (modules) from the abnormal human brain (Figure 6.2) distribute – as communities - into five distinct clusters. Note that a dot represents a mathematical marker connected (dark blue line) either to a duplicate marker or to a disorder (From Bolender, 2015).

Step 2: Next, the complexity of each cluster identified in Figure 6.3 is unfolded to reveal the next level

of complexity (Figure 6.4). Now clusters labeled 1, 2, 4, and 5 in Step 1 display clusters of their own. Note that cluster 3 – shown in Figures 6.3 and 6.7 - contains a single disorder (Alzheimer's disease).

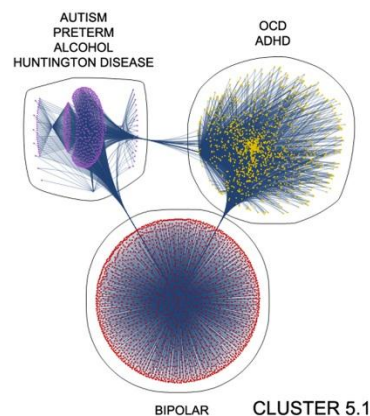
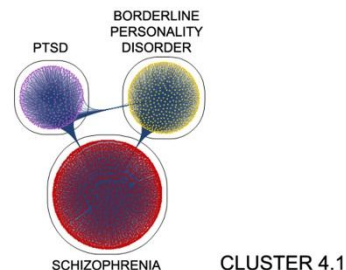
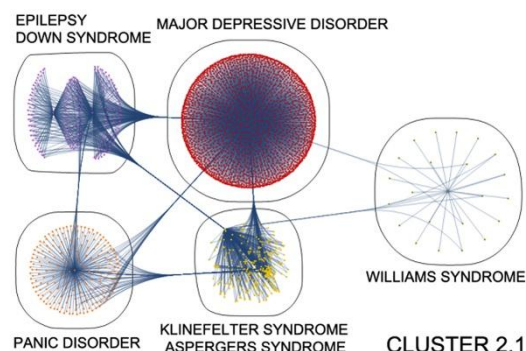
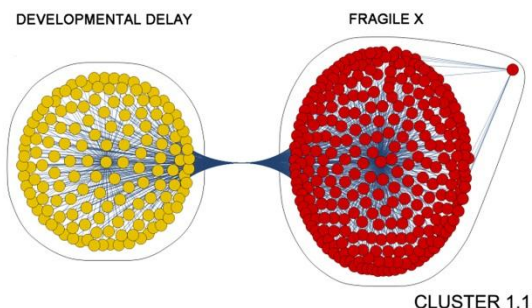


Figure 6.4 Step 2: The CommunityGraphPlots illustrate the sharing of markers (modules) between disorders. Each dot represents a mathematical marker and the blue line can be a connection between two shared markers or between a marker and a disorder. Each cluster is characterized by the disorder(s) it contains (From Bolender, 2015).

Step 3: Whenever a given cluster in Figure 6.4 carries more than two disorders, it is unfolded further (Figure 6.5).

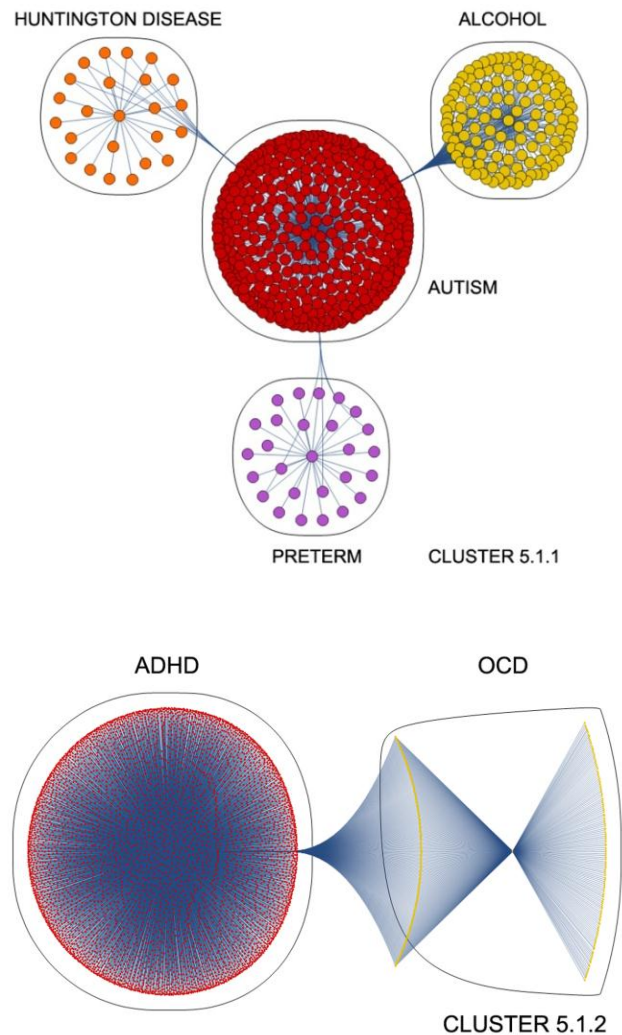
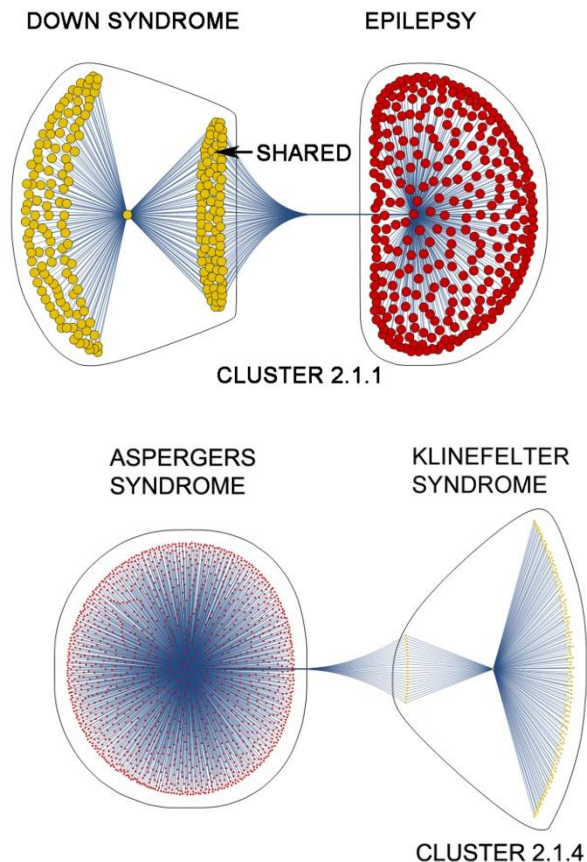
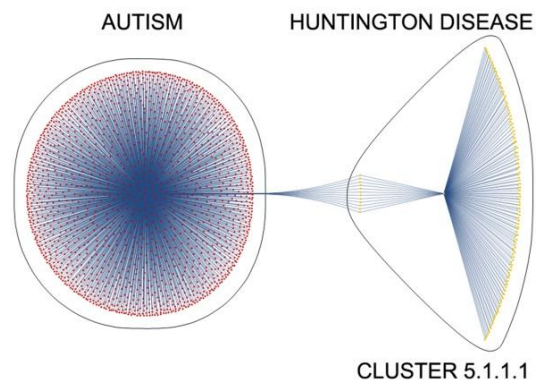


Figure 6.5 Step 3: Clusters containing multiple disorders in Step 2 are unfolded into clusters containing just two disorders. In such clusters, the shared markers (modules) appear as an intermediate, spindle shaped structure (From Bolender, 2015).

Step 4: Finally, cluster 5.1.1 is unfolded into three pairs of clusters (Figure 6.6).



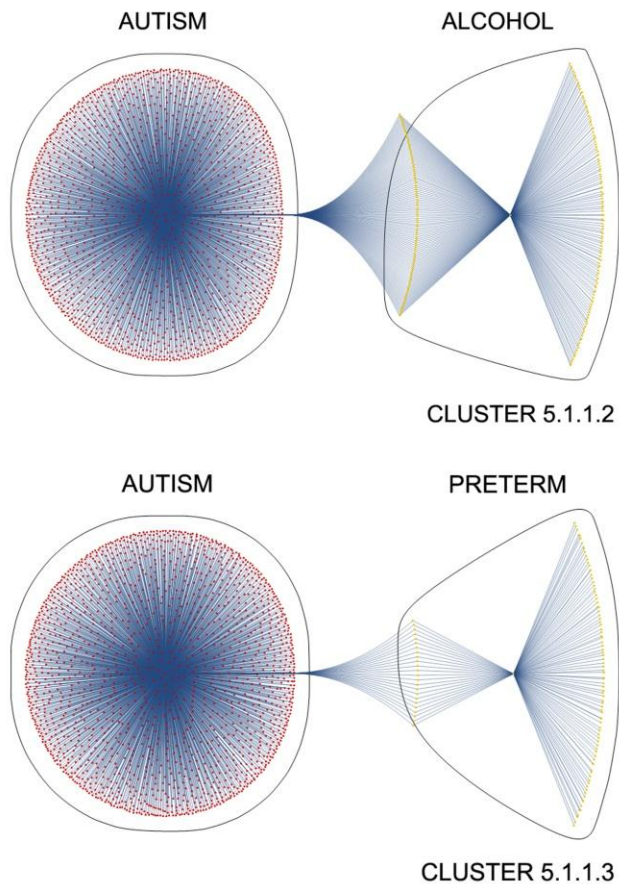


Figure 6.6 Step 4: Cluster 5.1.1 is resolved into three clusters relating the markers (modules) of autism to Huntington disease, alcohol, and preterm (From Bolender, 2015).

Figure 6.7 summarizes the disorders in the five clusters of Step 1 (Figure 6.3). Each cluster identifies disorders most closely related – as defined by the sharing of markers. Bear in mind, however, that this picture reflects the contents of the current IBVD database and is likely to change over time. Whether protocols developed to treat a specific disorder might also benefit disorders in the same cluster becomes a question we are encouraged to ask.

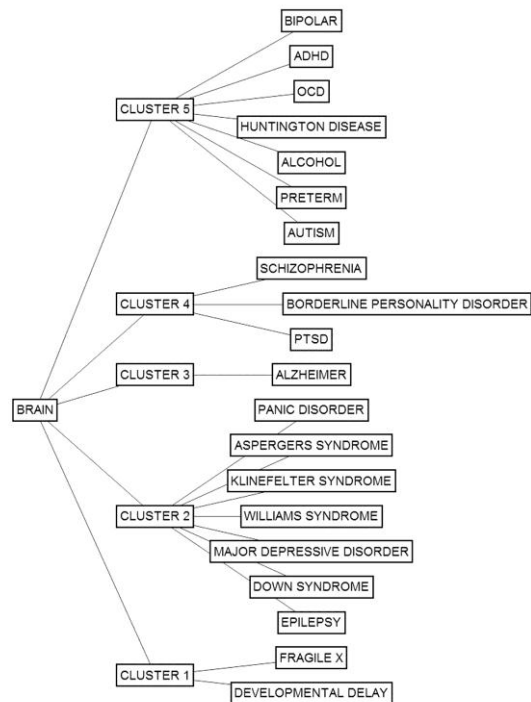


Figure 6.7 The graph shows the relationship of clusters to disorders in the composite human brain, unfolded as a function of shared markers (Figure 6.3; From Bolender, 2015).

6-4 Sharing Markers

When reduced to the final cluster pair, we can identify the specific markers being shared by the two disorders. Figure 6.8, for example, adds detail to the relationship of ADHD to OCD by replacing the dots of Figure 6.5 with the alphanumeric strings of mathematical markers. Notice that the OCD cluster shares more than half (58%) of its duplicate markers with those of the ADHD cluster. Given the extensive sharing of markers (Figure 6.3, Table 6.2), it appears likely that substantial fractions of many disorders will map back to the genome with similar routes and destinations. If this turns out to be the case, then identifying, targeting, and disrupting the most damaging routes may prove to be an effective strategy in managing groups of related disorders.

Recall that a given disorder carries a complement of shared and unique markers defined by biology and to an unknown extent by the current contents of our parallel complexity.

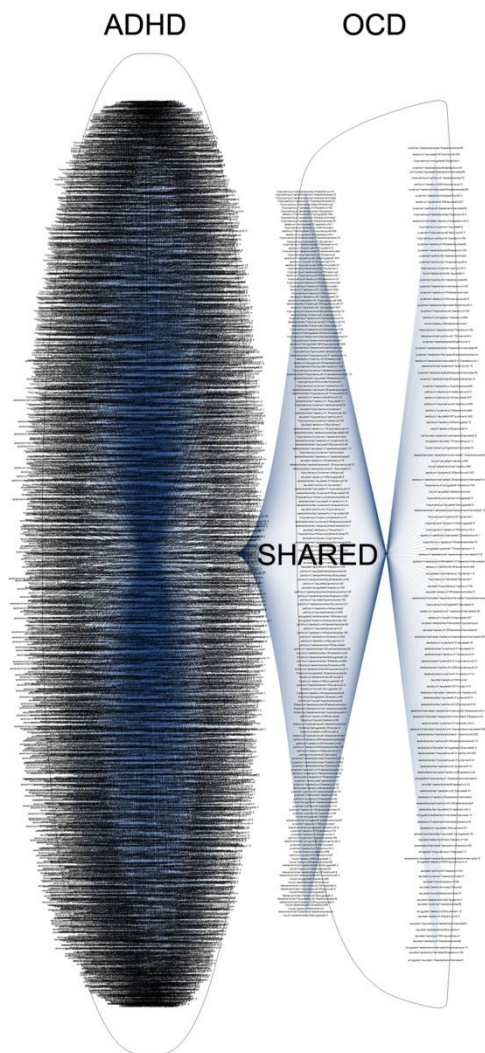


Figure 6.8 The clusters formed by ADHD and OCD show a strong sharing of markers. Notice that a similar pattern of extensive sharing exists between autism and alcohol (Figure 6.6; From Bolender, 2015).

Move 27: Can we unfold the complexity of brain disorders into well-defined communities of markers and disorders?

Yes, the mathematical markers associate preferentially with different disorders, thereby forming communities (clusters).

Move 28: Can disorders sharing similar markers share similar symptoms?

6-5 Sharing Markers and Symptoms

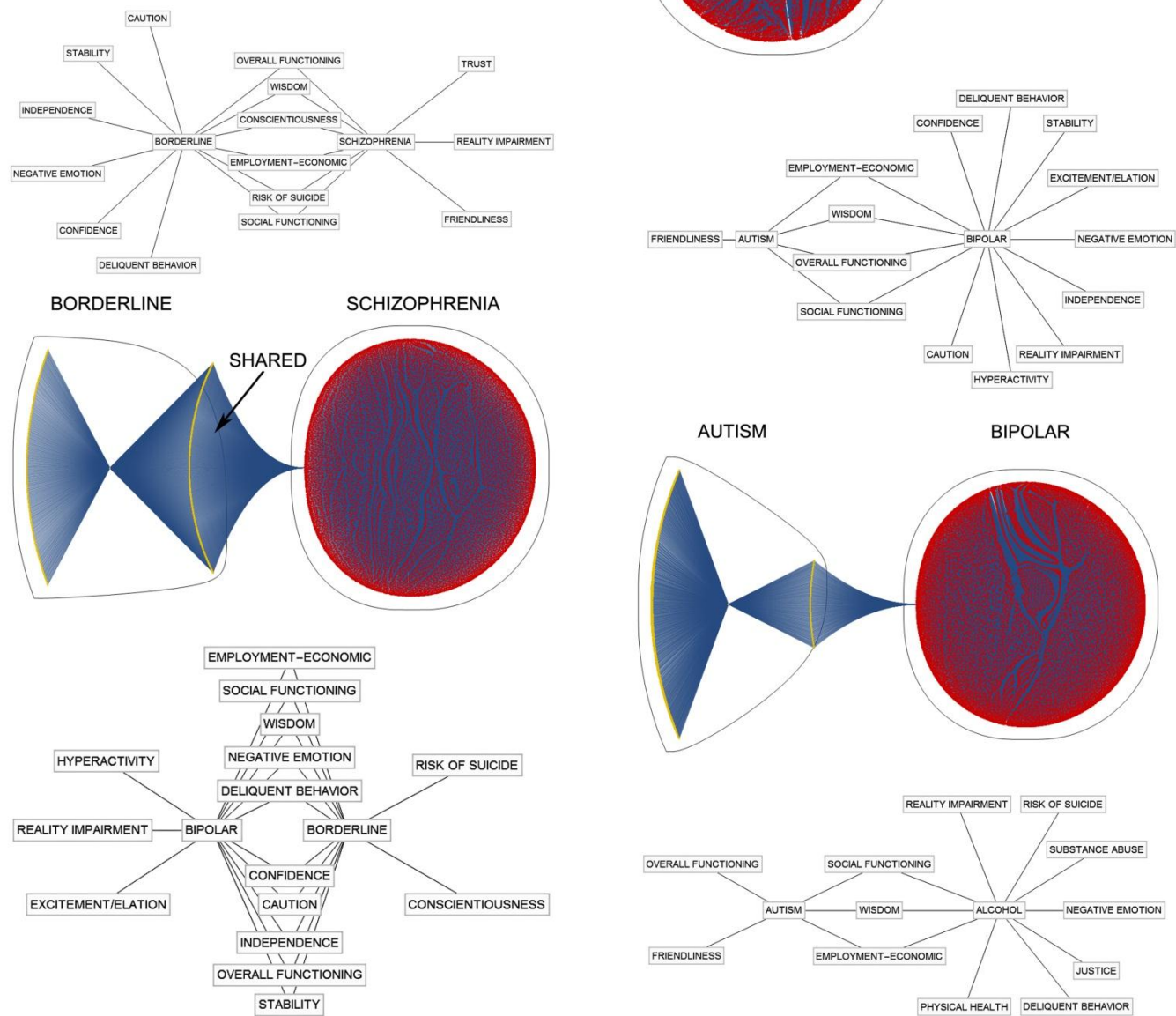
A clinical diagnosis seeks to determine the nature of a disease or disorder, typically by identifying symptoms. When diseases share similar symptoms, which is the case for many disorders of the brain, a differential diagnosis becomes the method of choice. Although Table 6.2 includes only a small sample of disorders and symptoms, it serves to illustrate the challenge faced by a physician when diagnosing a disorder of the brain.

Table 6.2 The table identifies symptoms for various disorders as impairments. Given the subjective nature of identifying impairments and the fact that a given impairment applies to many different disorders, making a differential diagnosis requires vast expertise (Adapted from Internet Mental Health © 1995-2015 Phillip W. Long, M.D., From Bolender, 2015).

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 EMOTION		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					
CONSCIENTIOUSNESS							X		X			X
CONFIDENCE						X	X	X				
INDEPENDENCE			X			X	X			X		
OVERALL FUNCTIONING	X			X	X	X	X	X	X	X	X	X

In this move, we will combine the information in Table 6.2 with the results of CommunityGraphPlots to look for correlations of symptoms to markers.

Figure 6.9 illustrates that disorders sharing similar symptoms frequently share similar markers. Although no attempt was made to associate symptoms to specific markers, larger data sets will encourage such studies.



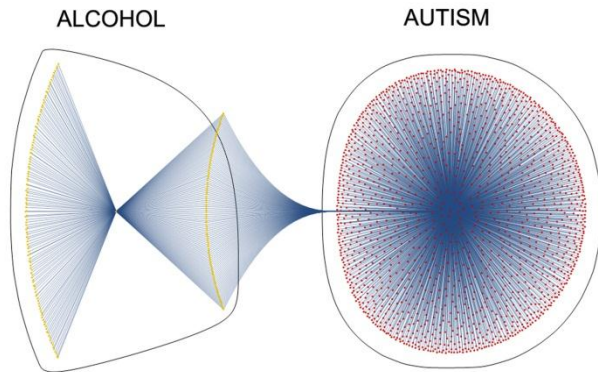


Figure 6.9 Graph and community plots illustrate the extent to which different disorders share similar symptoms and markers (From Bolender, 2015).

Move 28: Can disorders sharing similar markers share similar symptoms?

Yes, preliminary results suggest that they can.

Next, we can unfold the composite brain of Figure 6.2 according to the number of times a given marker is duplicated. This will allow us to identify those markers – with their associated parts and ratios – most often associated with the disease process. Such information might provide clues about where the disorder started or what parts must be in play for the disorder to exist. By plotting duplicates (Figure 6.1) ranging from 2 to 64, we can identify – at least provisionally - the relative importance of specific parts. This becomes our next move.

Move 29: Can we rank order the relative importance of markers in a community as a function of their frequency?

6-6 Identifying the Prime Movers

Figures 6.10 to 6.17 display CommunityGraphPlots for markers and disorders (above) and for parts and disorders (below) for duplicates ranging from >11 to >2 per group. They are useful in that they identify the preferences shown by disorders for specific

markers, parts, and connections. The figure legends identify the parts, the names of which are often obscured in the clusters.

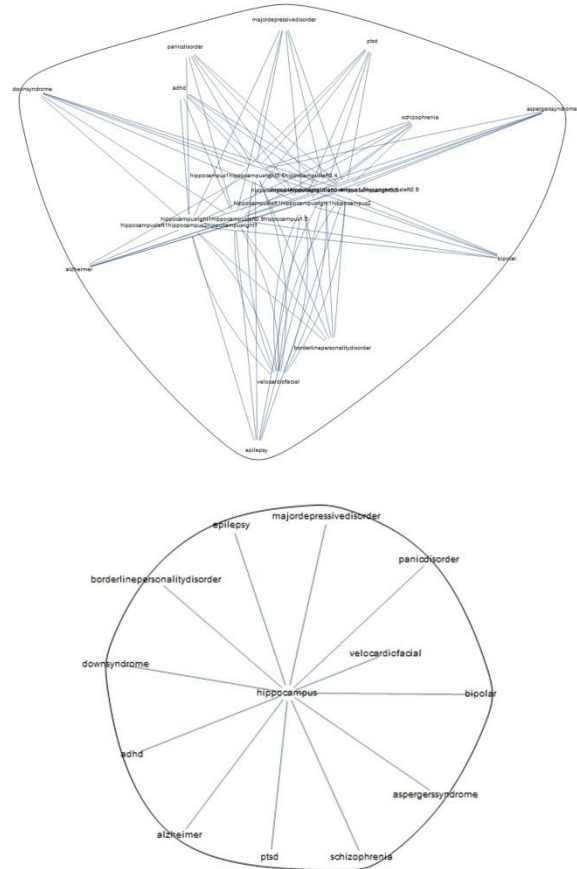
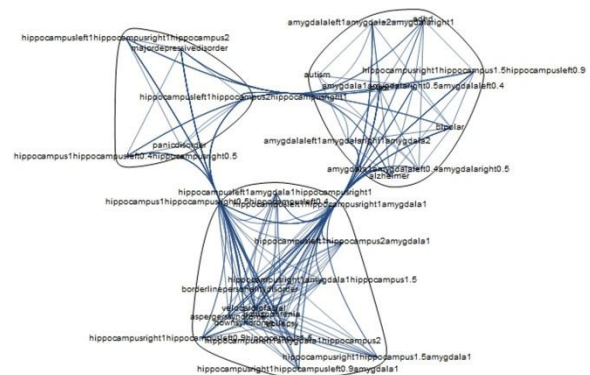


Figure 6.10 Duplicates >11. The hippocampus is the part of the brain involved most often in the disease process (From Bolender, 2015). Note that some of the images can be enlarged to read the fine print.



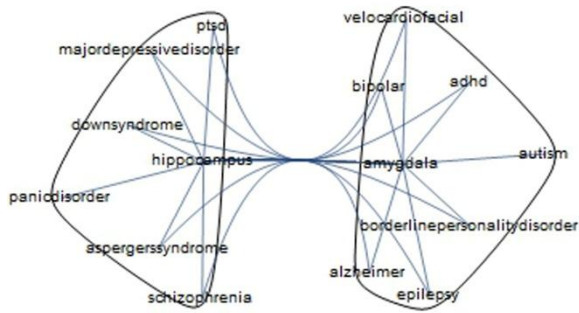


Figure 6.11 Duplicates >9. The disorders cluster around the hippocampus and amygdala (From Bolender, 2015).

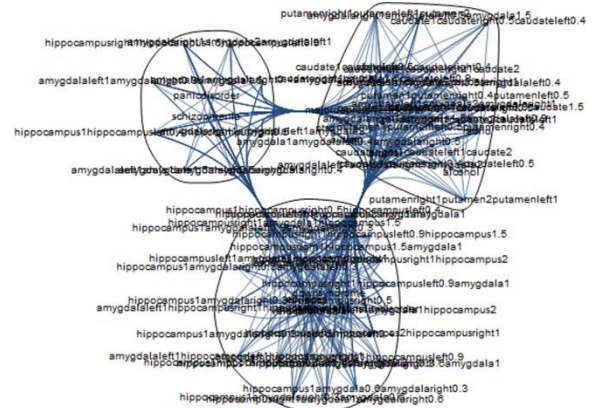


Figure 6.13 Duplicates >7. The disorders cluster around the putamen, caudate, and hippocampus-amygdala (From Bolender, 2015).

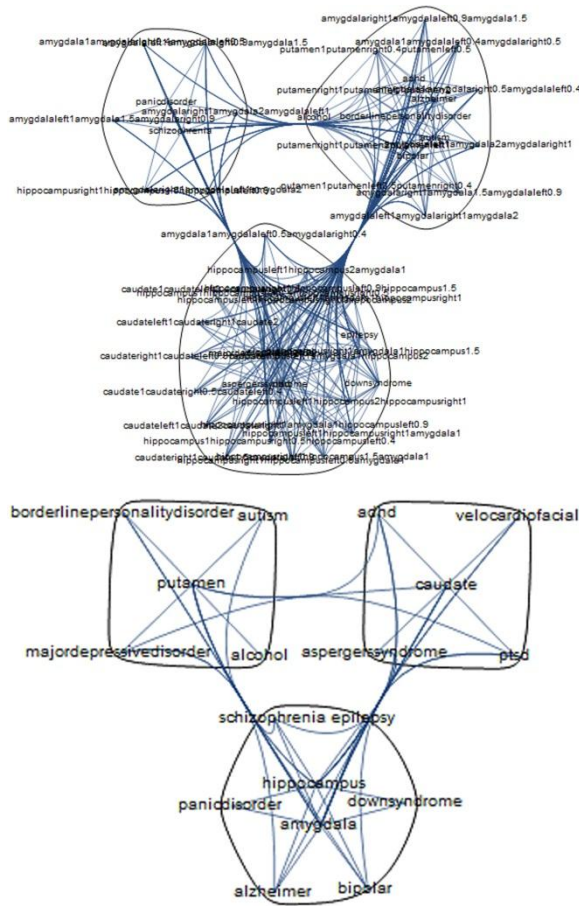


Figure 6.12 Duplicates >8. The disorders cluster around the putamen, caudate, and hippocampus-amygdala (From Bolender, 2015).

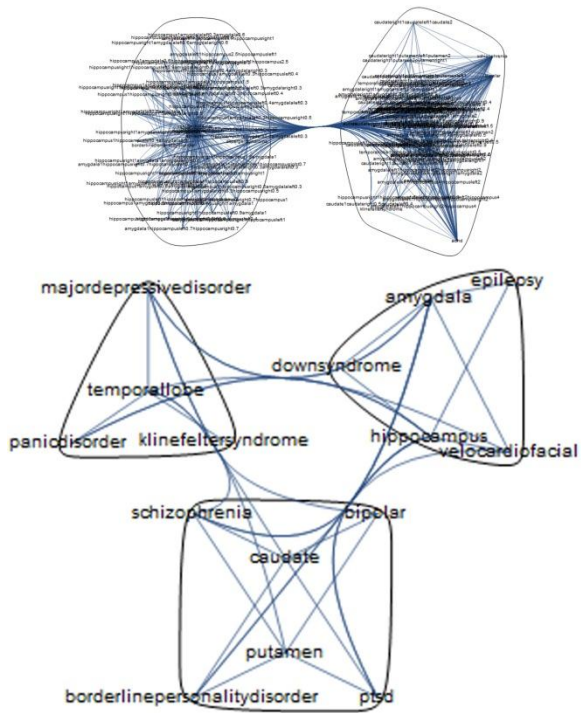


Figure 6.14 Duplicates >6. The disorders cluster around the lateral ventricle – brain, cerebrum-palladium, hippocampus-temporal lobe, and putamen-caudate- thalamus (From Bolender, 2015).

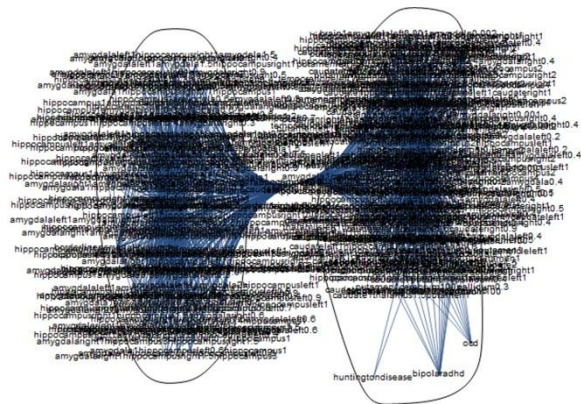


Figure 6.15 Duplicates >5. The disorders cluster around the thalamus-nucleus accumbens, anterior and posterior insula, brain-amygdala-cerebrum-caudate-putamen-pallidium (From Bolender, 2015).

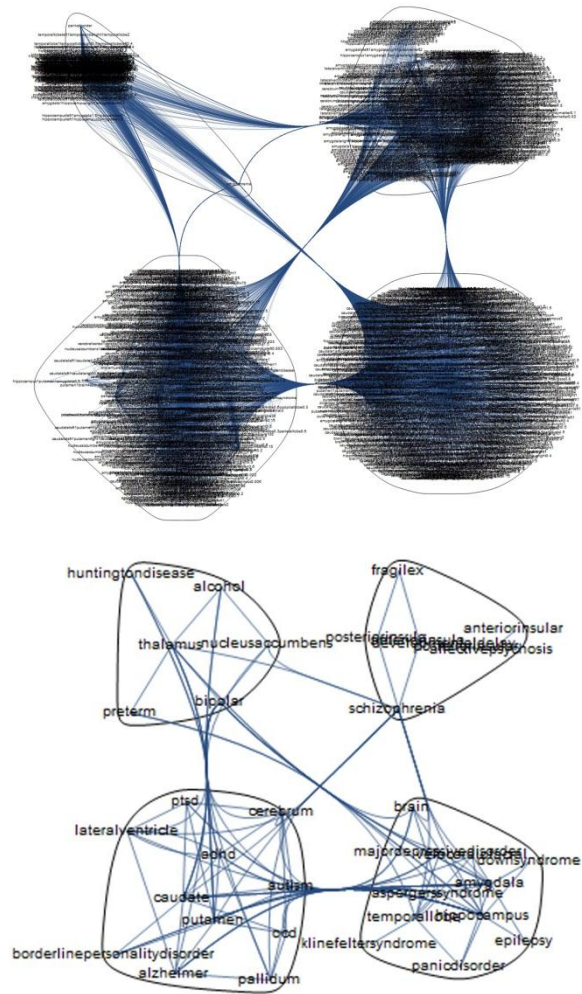


Figure 6.16 Duplicates >3. The disorders cluster around the brain-thalamus, anterior and posterior insula, nucleus accu-

bens-cerebrum-pallidum-caudate-putamen-lateral ventricle, and hippocampus-amygdala-temporal lobe (From Bolender, 2015).

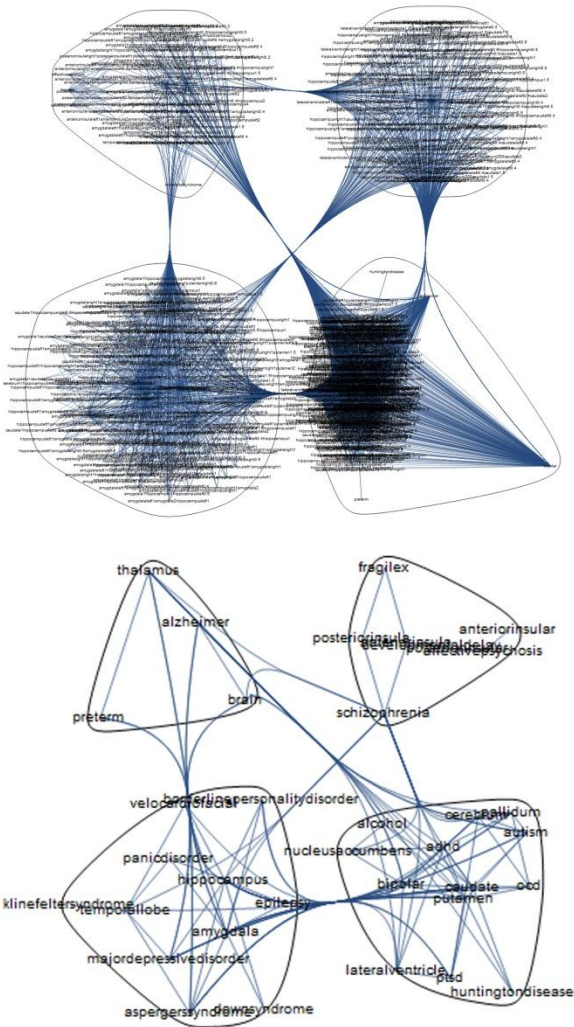


Figure 6.17 Duplicates >2 (From Bolender, 2015).

Table 6.3 summarizes the relationship of parts to disorders according to the frequency of duplicates (>11 to >2) – illustrated in Figures 6.10 to 6.17.

Table 6.3 The table shows the relationship of parts to disorders arranged according to the number of duplicate markers. These data suggest that a relatively small number of parts play a disproportionately large role in the disease process (From Bolender, 2015). Recall that we found a similar pattern earlier with the biological blueprint (Figures 2.16 and 2.17).

	>11	>9	>8		>7		>6		>5			
	HIPPOCAMPUS	AMYGDALA	PUTAMEN	CAUDATE	PUTAMEN	CAUDATE	PUTAMEN	CAUDATE	PUTAMEN	CAUDATE	PUTAMEN	CAUDATE
adhd	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		
down-syndrome	X	X		X		X		X		X		
epilepsy	X		X	X		X		X		X		
huntington-disease												X
kliefelter-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		

Notice in Table 6.3 that most of the disorders depend importantly on abnormalities in just five parts – the hippocampus, amygdala, putamen, caudate, and temporal lobe – according to the data currently available in the IBVD. It also shows an extensive sharing of parts among different disorders. Taken together, Figures 6.10-6.17 and Table 6.3 continue to suggest a modular origin of disorders, wherein communities serve to identify preferences.

Move 29: Can we rank order the relative importance of markers in a community as a function of their frequency?

Yes, the community plots of duplicate markers can identify the parts most often involved in the disease process.

6-7 Summary of Chapter 6

The moves described in this chapter used a parallel complexity designed as a database of shared mathematical markers (modules) to unfold the complexity of the disease process. The results show that biology defines itself – in health and disease - as a structural hierarchy using modules consisting of clearly defined parts and connections. Since the same modules frequently appear both locally and globally, we satisfy the reproducibility and validity requirements of complexity theory and offer convincing evidence that the brain routinely uses many of the same modules to assemble different disorders.

Two things bear mentioning. It takes a large number of different modules to produce a disorder and all disorders draw many of their modules from what appears to be a common pool. In effect, a given abnormal module – like many genes - can produce more than one outcome. The challenge for biology in assembling a disorder is to get the right mix of markers and emergent properties to produce a given set of symptoms. Since a given disorder is reproduc-

ible within a population, the same or closely similar algorithm must be conserved and in play to produce a recognizable phenotype. This being the case, it should be possible to unfold the complexity of a disorder back to the genome and in so doing reconstruct the causative algorithm.

Move 29 included a summary of the disease process based on community graphs (6.10-6.17) and the frequencies of shared markers (Table 6.3). If we equate frequency of occurrence to its importance in the disease process, then the hippocampus emerges as the most influential player (Figure 6.10). Does this mean that disorders of the brain may not appear in the absence of an abnormal hippocampus or that an abnormal hippocampus must exist to trigger the disorder of other parts, such as the amygdala, caudate, and putamen? Although such questions remain beyond the reach of the current databases, our recent progress in working out a quantitative relationship of diagnosis to prediction (Chapter 5) becomes increasingly relevant.

Chapter 7

Caveats

Biology explores life as a complexity, whereas we – as scientists – prefer to pursue it as a simplicity. This schism between reality and what we choose to perceive as reality provides fertile ground for cultivating caveats.

A caveat is a warning. It scratches beneath the surface to uncover hidden dangers, ambiguities, and half-truths. As soon as we leave simplicity and take even a single step toward complexity, many of our current perceptions face challenges. What we once imagined to be possible often becomes impossible and impossible possible. This can force us, for example, into uncomfortable positions by questioning what we perceive to be true. In short, caveats have the nasty habit of looking at well-established “truths” and seeing reckless assumptions.

7-1 Change

Assumption: Change is simple, not complex.

If we are not measuring or estimating a volume, surface, length, or number of particles directly, detecting a biological change can become problematic. When collecting biological data as concentrations and using them directly to detect changes, we are assuming that they behave as simplicities. If true, then both concentrations (complex) and absolute values (simple) would always give the same results. This, of course, is not the case (Chapter 1). The pay-back for making this assumption is a literature polluted with bias, misinformation, and contradiction.

7-2 Stereology

Assumption: Stereology can eliminate the ambiguity of concentrations by evaluating hierarchy equations.

In theory, yes, in practice often no. Hierarchy equations carry the assumption that all the reference spaces in play – the denominators of the concentrations – will cancel even when they carry very differ-

ent distortions (biases). This assumption, when put to the test, carries a substantial risk (Chapter 4).

7-3 Data Points

Assumption: The best way to detect a biological change is to divide an experimental data point by that of a control.

In effect, this universally accepted approach to studying biology is seriously flawed. Change in biology involves a forest, not just a few trees. Given the connectivity of its parts and the rules in play, biology creates a highly dispersed pattern of change, rather than changing just a few of its parts in isolation. Assuming that we can interpret a change by following the behavior of one or a few parts misses the meaning of a biological change altogether (Chapters 3-6). In a complexity, change exists as a pattern wherein biology creates new relationships of parts to connections.

7-4 Data Equivalency

Assumption: Data collected from living and nonliving sources are compatible.

In view of the findings presented in Chapter 4, this assumption appears indefensible. Quantitative data collected from living and nonliving sources were found to be largely incompatible. Although the results presented in Chapter 4 require independent confirmation, they come from the best data currently available – refereed research produced by experts.

A key insight to come from Chapter 4 is that a gold standard coming from living individuals allows us to minimize the effect of biases produced post-mortem. Such a standard provides corrections that can convert incompatible data sets into compatible ones (Figure 4.11). By making data derived from living and nonliving sources interchangeable, we put a large portion of the biology literature back in play (Chapter 4). This becomes an important issue be-

cause extracting patterns on our way to the genome will rely heavily – by necessity - on post-mortem data obtained from light and electron microscopy with stereological methods.

7-5 Volume Independent Methods

Assumption: Detecting changes reliably in post-mortem biology with stereology is limited to the volume independent method of counting – the fractionator (Gundersen et al., 1988).

This statement is true, but only as long as the volume dependent estimates (V , S , L , N , V/V , S/V , L/V , and N/V) continue to carry the bias of the volume distortion. By switching to patterns based on ratios, multiple solutions to this problem of volume dependency can be identified and applied (Chapter 4; Bolender, 2013).

7-6 Reproducibility

Assumption: Reproducibility of results at the global level cannot be expected to occur routinely with biological data.

In the absence of reproducibility, biological data become subject to disbelief. Moreover, such a deficiency keeps us in a weak position. When multiple outcomes exist for the same experiment, we must decide which one to believe. This pushes us into making choices that are often difficult to defend.

We know that detecting biological differences requires sample sizes sufficiently large to overcome the variation of a population and that such variation is influenced by random (imprecision) and systematic (bias) errors. We also know that biases can change and that variation in biology exists within and between subjects.

If, instead, we start with the results and identify a data set capable of delivering widespread reproducibility, then it would appear that we have tapped into an information source wherein the major sources of variation are minimized. Our databases become such a source when the same mathematical markers appear repeatedly at a global level. Since such duplicate markers occur in the tens of thousands, reproducibility appears to be an inherent

property of biology. As such, mathematical markers represent a highly reproducible form of biological data (Chapters 4-6).

Mathematical markers also pass notably harder tests of reproducibility. Usually, measures of variance involve only two variables (one data point for the control and another for the experimental), whereas mathematical markers routinely demonstrate reproducibility with substantial numbers of duplicate ratios comparing two (data pairs), three (triplets), and four (quadruplets) numerical variables at a time. Such results offer evidence at a global level that biology is running a very tight ship, wherein strict rules of connectivity exist and are being obeyed.

7-7 Biological Variation

Assumption: Biological variation severely limits our ability to detect small differences in the human brain.

The IBVD includes – online - an extensive collection of scatter plots summarizing data collected from various parts of the brain. They demonstrate – with upmost clarity - that the volume of the same part can vary enormously from one individual to the next. In fact, the data points create amorphous clouds of points with little hint of order. Detecting significant differences between such diffuse clouds of data often becomes a largely hopeless exercise. These scatter plots remind us that biology gives the brain considerable leeway in making an individual part, a pattern we choose to identify as biological variation.

Our response to such variation is to design experiments in ways that allow us to detect significant differences by relying on a variety of statistical approaches. However, a curious uneasiness surrounds this type of solution. Something appears amiss. Why would biology - so committed to optimizing outcomes - tolerate such apparent disorder? Is it possible that the biological variation associated with these isolated data sets is merely a construct of our own imagination? If our isolated data display generous amounts of biological variation, will biology – when we address it as a complexity – go along with our definition of its variation?

A following worked example shows how we can use complexity theory to answer such questions. Going from reductionism to complexity is like going from chaos to order. We can see this transformation occur by plotting data taken from the IBVD. Figure 7.1 illustrates 61 estimates (control and experimental) for the volume of the amygdala, sorted according to size. The chaos appears in the figure as a broad range in the individual values plus the expected discontinuity produced by the data carrying different units.

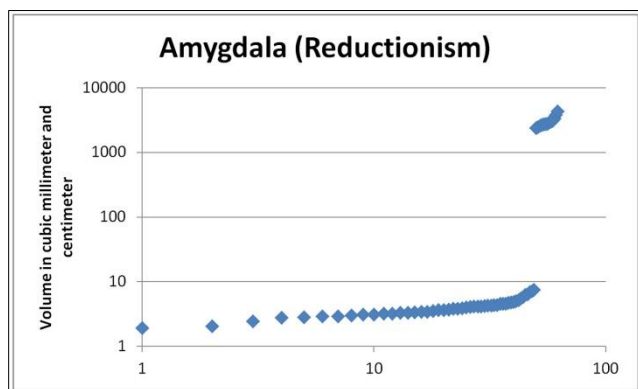


Figure 7.1 MRI estimates of volume for the human amygdala display a pronounced biological variation. In the absence of connections, the isolated data become chaotic. Notice that the plot is log-log (From Bolender, 2012).

If we add back the connections, by creating ratios between the left and right parts of the amygdala, the intrinsic order of the complexity begins to reappear (Figure 7.2).

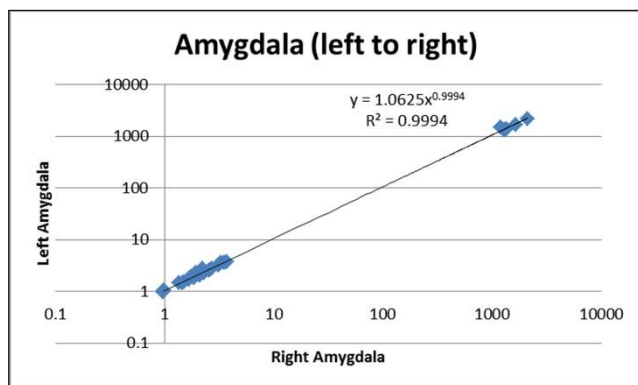


Figure 7.2 When expressed as ratios, the MRI data of Figure 7.1 display a high degree of order, as detected by a power regression with an $R^2=0.9994$ (From Bolender, 2012). Notice that the equation is almost linear (1.0) with its slope of 0.9994.

Next, we can calculate a triplet ratio (X:Y:Z) wherein X is set equal to one. This allows us to express the Y, Z data as a repertoire equation (Figure 7.3).

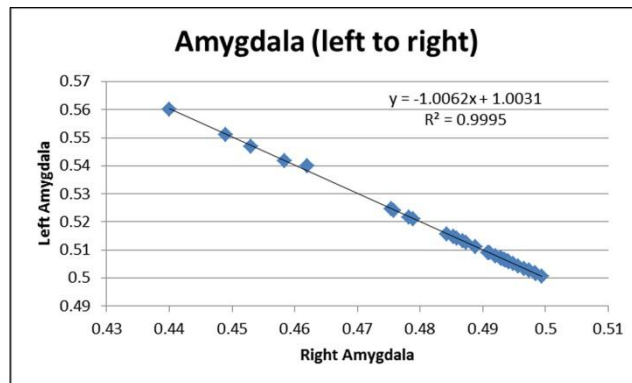


Figure 7.3 After forming the ratio X:Y:Z where X=1, Y (left) is plotted against Z (right). The result is a linear equation with an $R^2=0.9995$ (From Bolender, 2012). Were biology using an equation to define the relationship of the left amygdala to right, it might be using this one.

Notice in Figure 7.3 that the range of the data has been compressed by orders of magnitude (compared to Figure 7.1) and that the two parts approximate linearity ($R^2=0.9995$). Most of the data now differ by less than five percentage points.

If Figure 7.1 represents chaos and Figure 7.3 order, then we want to be somewhere in between near the edge of chaos, the place where the most interesting things happen (Walthrop, 1992; Kauffman, 1995). This final step consists of going from repertoire values (Figure 7.3) to decimal repertoire values (Figure 7.4). Notice what happens. All the data shown in Figure 7.1 condense into a single ratio (0.4:0.5).

The example illustrates how complexity theory can optimize the effectiveness of our data. The original reductionist data displayed biological variation on a grand scale (Figure 7.1), whereas the same data viewed in a complex setting detected mathematical patterns (Figures 7.2-7.3) with little if any biological variation. Finally, optimizing the data globally identifies an underlying design principle of biology (Figure 7.4). The amygdala of the human brain in health and

disease exists left to right in the ratio of 1:1.25; it represents a biological rule. By knowing such rules, we have a far better chance of figuring out where they come from.

This set of figures suggest that biological variation severely limits our ability to detect small differences in the human brain – but only when we agree to accept the chaos created by reductionism and by a statistical construct of our own making. In contrast, biology, cannot afford such extravagant behavior. Biological variation certainly exists, but not in the bloated amounts we choose to assign to our data.

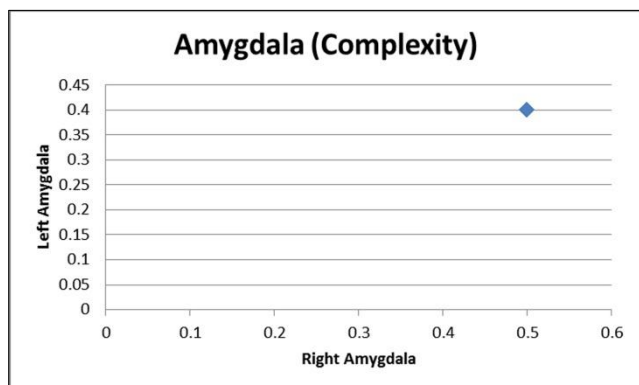


Figure 7.4 A single decimal repertoire value (ratio) for the amygdala (Y:Z) - left vs. right – summarizes the amygdala data across many different publications and diseases. Notice that all 61 data points are now represented by a single ratio (Y:Z = 0.4:0.5). Multiple copies of this mathematical marker - amygdala1amygdalaleft0.4amygdalaright0.5 – exist in the MRI database (From Bolender, 2012).

7-8 Experiments

Assumption: The information derived from a single experiment provides enough information to make meaningful interpretations.

Since a biological change triggers multiple events, following the behavior of just a few variables fails to capture the global nature of change. Interpreting a biological change as a complexity is an operation better suited to large-scale databases where data can be interpreted by identifying analogous patterns and cross correlations. Within the space defined by a parallel complexity, for example, we interpret what we do not know within the framework of what we do know. In such a setting, databases become cen-

tral to our task of setting a scientific argument and defending its conclusions.

7-9 Modifying the Human Genome

Assumption: The human genome can be modified responsibly in the absence of detailed phenotypic information.

In a complex system, where small perturbations can result in large effects and produce unintended consequences, this proposition is likely to get us into deep trouble. Since our manipulation of genomes is already well underway in plants and animals, tracking and interpreting the consequences at the level of the phenotype would seem to be an essential component of the process. Complexity in biology is the product of a finely tuned machine that may or may not respond positively to our reprogramming efforts. Ignoring this complexity ignores the certainty of unintended consequences.

7-10 Published Research Findings

Assumption: The biomedical sciences enjoy a distinguished record of success.

“Why most published research finding are false.” Such is the title of a now famous paper by J. P. A. Ioannidis (2005). He maintains that research findings may be little more than accurate measures of the prevailing bias. Moreover, at a recent Peer Review Congress, he noted that a meta-meta-analysis related to cancer research shows that the scientific studies are “correct” in almost no cases. He also reported that linking genes with particular diseases are correct only 1.1% of the time and noted that both biomarkers and prediction models for diseases have dismal records of success. In an analysis of bias in 17 million papers, he detected 235 sources of bias. After raising a red flag, he continues to wave it vigorously. Corporations have also become wary of published data. Amgen, for example, found that 80-90% of preclinical studies could not be replicated.

In a recent study, Colquhoun (2014) points out that accepting a $P=0.05$ leads to in incorrect result at least 30% of the time. Moreover, he suggests that underpowered experiments will be wrong most of

the time. In fact, our time honored tradition of accepting a significant difference at $p \leq 0.05$ appears well on its way to becoming a relic of a bygone era. A recent editorial published in Basic and Applied Social Psychology (2015) declared the null hypothesis significance test to be 'invalid' and has banned it from future papers submitted to the journal.

Statisticians appear to want reproducibility supported by independent validation. This requires substantial reductions in methodological bias and biological variation, goals largely inconsistent given our current theory structure. We have witnessed throughout this book the mischief being created by these uncontrolled sources of noise. Comparing biological concentrations produces incorrect results about 50% of the time (Chapter 1), ignoring valences distorts results (Chapter 2), absolute estimates carry volume distortions (Chapter 4), post-mortem estimates do not correspond to those coming from their living equivalents (Chapter 4), biological variation is allowed to run rampant (Chapter 7), and "...most published research findings are false." (Chapter 7).

7-11 Biology as a Science

Assumption: Biology, which has evolved into a descriptive science, cannot be expected to address problems at a fundamental level.

Such a statement is true for a theory structure based on reductionism, but not for one based on complexity. The responses to the caveats listed above derive from a complex biology operating on a solid, mathematical foundation.

7-12 Summary of Chapter 7

The point in listing caveats is to call attention to the present state of our science. Currently, we are failing miserably on a number of fronts because we tenaciously hold to the belief that we can study a complexity as a simplicity. If parts and connections define a complexity and we throw away the connections, then the complexity ceases to exist. In such an artificial setting, we can study biological parts, but we cannot study biology. Curiously, we remain unaware of the fact that biology can exist as a science only when it exists as a complexity.

In addressing these caveats, our best option is to reinvent biology as a quantitative discipline operating within a theory structure based on complexity. Although this approach may seem difficult and intimidating at first, it actually makes our job much easier and far more effective.

Chapter 8

Theory of Biological Complexity

8.1 Introduction

The overarching principle of the new theory is that it takes a complexity to solve a complexity. This means that to test the theory empirically we need to construct a parallel complexity as close to the original as possible, relying exclusively on the rules that exist first in biology and then mirrored in our reconstituted complexity. The design-based sampling methods of stereology play an important role in this building process by providing unbiased sampling and offering access to parts of all sizes.

Complexity is an unfamiliar place. New rules apply, our perceptions change, and we ask and answer questions differently. The first order of business is to learn the rules of the game, which in science consists of generating a new theory structure. This represents an ongoing process wherein the theory evolves in step with the discovery process.

Recall that the fundamental building blocks of a biological complexity include parts and connections. Volumes, surfaces, lengths, or numbers define the parts quantitatively and ratios derived therefrom the connections. From this simple beginning, the complexity of an organism grows as the parts and connections cascade throughout the hierarchical levels of an organism. Since everything consists of the same basic building blocks and all the blocks are connected, our parallel complexity begins to resemble the original biology – at least on a modest scale. Testing the theory consists of looking for persistent patterns - locally and globally – and then using these patterns to define the rules of the game.

A collection of working lists, including goals, requirements, basic principles, definitions, derivatives, first principles, and rationale summarize the progress to date in constructing this new theory structure.

8.2 A First Principles Approach

Biology – as a science – can be said to derive from first principles if it relies on the basic and established laws of nature. In the absence of such principles, biology defaults to the models and assumptions driven largely by convenience.

Can we derive biology from first principles? The answer is yes. When we bring the power of data coming from thousands of highly skilled investigators into conjunction, the basic rules and laws of biology appear almost effortlessly. This is the story told by the preceding chapters.

The argument for a first principles approach becomes even more compelling because of the opportunities it creates. If the biology enterprise – on a scale of one to ten – is currently winning at a level of one or two, what would happen if we moved it up a notch? Innovation and discovery would begin to explode. 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.”*

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 mathe-

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 with post-mortem data, and unfold the disease process in the human brain.

8.3 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 patterns (e.g., mathematical markers) capable of scaling – by rule - at both local and global levels. Typically, biology displays its complexity as a stoichiometry based on the ratios of its parts. It applies this rule to create both order and disorder. For our purposes here, we define a rule as a mathematical pattern, one that exists simultaneously at local and global levels.

8.4 Theory Structure

The theory structure includes a current set of guidelines for exploring biology as a complexity. They derive from the published data of refereed publications – numbering in the thousands. The following list summarizes the goals evolving with the theory structure.

8.4.1 Goals

- Generalize the data of the biology literature.
- Define and assemble a data-driven approach to the basic and clinical sciences.
- Identify mathematical patterns in biology.
- Explore biology as a rule-based system.
- Use published data to create a parallel complexity based on rules intrinsic to biology.
- Minimize bias in experimental systems.
- Minimize biological variation.
- Maximize reproducibility.
- Remove post-mortem distortions by harmonizing pre and post-mortem data.

- Demonstrate with practical examples the effectiveness of a new approach to problem solving based on empirical data and guided by the rules of biology.
- Capture biological phenotypes mathematically and use them to diagnose and predict outcomes.
- Evaluate current methods of collecting and interpreting data in the basic and clinical sciences.
- Assemble diagnostic databases from the biology literature capable of diagnosing disorders of the brain correctly - 100% of the time.
- Scale the application of mathematical markers from small data sets to large.
- Develop methods for extracting meaningful patterns from large data sets.
- Identify algorithms and strategies used by biology to create disorders of the brain.
- Develop and distribute software and databases that can accelerate productivity by leveraging published data into problem-solving tools.
- Develop a strategy for connecting phenotypes to genotypes.
- Optimize outcomes.

8.4.2 Data Requirements

- Collect biological data with unbiased sampling methods.
- Express data as volumes, surfaces, length, or numbers. Concentration data formed from these and other parameters are subject to specific rules and limitations (See earlier reports; Bolender 2001-2015). Note: Most data types in biology readily meet this requirement, although digging into the units may be required.
- Assemble data as connected sets, consisting of ratios, mathematical markers, strings, networks, et cetera.
- Integrate data quantitatively within and across hierarchical levels.
- Use a common format – based on ratios - to organize and generalize data.

- Configure data so that they can detect the same pattern locally and globally.
- Operate within the bounds of a complexity parallel to the one defined by biology.
- Correct the volume distortions associated with post-mortem data.
- Tune data sets – by applying filters - to enable diagnostic and predictive properties.
- Store and distribute data in digital form.
- Encourage open access to published data in multiple formats.

8.4.3 Basic Principles and Definitions

- A biological complexity consists of parts and connections distributed hierarchically.
- Complexities can be local, global, and nested.
- A biological complexity can unfold into smaller patterns or fold into larger ones.
- Parts and connections define the organizational framework of biology as distinct and quantitative patterns. As such, they represent a rule-based system.
- A parallel complexity represents a data-driven construct designed specifically to emulate biological complexity.
- Ratios and derivatives thereof (i.e., mathematical markers) serve as the basic units of information and signal the presence of rules in a parallel complexity.
- Mathematical markers include parts (names) and connections (ratios) expressed as alphanumeric strings that can be snapped together to create compound strings.
- An artificial complexity, which exists in post-mortem data, is a product of the methods of specimen preparation and data collection.
- Parts display quantitative (volume, surface, length, number) and qualitative properties (names, locations).
- All parts are connected or connectable by forming ratios.
- A ratio defines the relationship of one part to another. Moreover, ratios define nested and modular sets of connections within and across hierarchical levels.

- Parts and connections form patterns that scale in size, beginning with a ratio of two parts and ending with a ratio of n parts - where n would include an entire organism.
- Patterns captured as mathematical markers increase their specificity as the number of parts and connections increase.
- In living subjects, mathematical markers routinely detect the same patterns (e.g., markers) locally and globally.
- In post-mortem subjects, mathematical markers infrequently detect the same local and global patterns, unless the data are corrected for volume distortions.
- Prediction in complex living systems involves interactions with parallel complexities capable of producing a correct diagnosis 100% of the time.
- Valances describe the ability of the same set of parts to form different numerical ratios (connections). They reflect biological rules of stoichiometry.
- Biological variation is a construct of its theory structure. Reductionism maximizes variability, whereas complexity theory minimizes it.
- In biology, change is defined by patterns.

8.4.4 Derivatives

A derivative includes - as a minimum - the names of two parts and their corresponding values formed into a ratio. In forming a ratio, the original published values may be used directly (repertoire value) or converted into a decimal step (decimal repertoire value). Data pair ratios take the form X:Y, data triplets X:Y:Z, and data quadruplets X:Y:Z:Q. Mathematical markers add the names of the parts to the ratios: AX:BY, AX:BY:CZ, and AX:BY:CZ:DQ.

Data Pairs, Triplets, and Quadruplets can be formed by inspection or by taking all possible permutations of the names of the parts – to which numerical values are subsequently assigned. When expressed as a decimal step (decimal repertoire value), the values combine with names to form mathematical markers. Such markers, which can use data before or after the application of corrections for the volume distortions

of post-mortem material, can display multiple valences.

8.4.5 First Principles (Rules)

To derive biology from first principles, we first need to know the principles. Such principles translate into biological algorithms that define biology as a complex adaptive system. They turn an information poor catalogue (genome) into an information rich masterpiece (phenotype).

Patterns that appear repeatedly identify biological rules.

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 forms strings, modules, and networks of 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 not in the relationship of one part to another. Typically, it maintains a given stoichiometric order, except when undergoing a change – e.g., growth, aging, and disease.

Rule 6: Biology defines complexity with modular structures divisible down to two parts with two values – the ratio (X:Y).

Rule 7: Biology consists of nested complexities, which can be unfolded and refolded mathematically.

Rule 8: Biology maintains patterns with redundant connectivity.

Rule 9: Biology defines change as a complex pattern.

Rule 10: Biology optimizes outcomes.

Rule 11: Biology grows in distinct steps, wherein patterns alternate between active growth (dynamic ratios) and inactive growth (stable ratios).

Rule 12: Biological parts can serve as dominant central organizers, wherein they form connections (ratios) with many other parts.

Chapter 9

Recommendations

9.1 Background

The marching orders for the Enterprise Biology Software Project came largely from the Biomatrix Group (Morowitz and Smith, 1987). We were charged with the task of organizing the published data of biology in such a way as to reveal connections, generalizations, and new theory structures. When we interpret published data as ratios, we can create a universal biology database capable of producing such outcomes in accord with the mathematical precepts of biological complexity.

9.2 Strategy

In this chapter, we extend the original directive of the Biomatrix by including additional recommendations. This upgrade seeks to establish the quantitative phenotype as a major player in the biology enterprise - a natural extension of an ability to interact with biology mathematically.

Recall that quantifying a phenotype required two databases. The first database, which came from the stereology literature, demonstrated that post-mortem data - when expressed as ratios - could detect biological rules locally. Detecting the same rules both locally and globally, however, required a second database derived from the MRI data of living subjects (IBVD). While stereology got us into the complexity game, we needed MRI to score. Both databases will be needed to win.

Using these two databases, we were able to upgrade our published research data by making the transition from one theory structure to another – from simple to complex. As a result, we have begun to redefine biology as a data driven, rule based, and quantitative science. By allowing relational databases to provide a universal framework for designing, analyzing, and interpreting experiments, we can effectively leverage the successes of the past and present to plan the future.

9.3 Issues

As a complex, information intensive science, biology requires unfettered access to large amounts of published data stored in relational databases. Herein lies a problem in that most of our research data continue to exist behind paywalls and not in relational databases. Our job – just as important as our research – will be to find an accommodation wherein the current win-lose situation can become win-win. Perhaps the easiest and most effective way of resolving this issue is to add a new requirement to the publishing process, one that includes entering data into public databases.

It would be a disservice to everyone to allow an adversarial relationship to develop between authors and publishers. Biology is clearly on a path to becoming an information science and everyone will benefit enormously by allowing this transition to occur.

9.4 Current Reality

Our current theory structure in biology is incomplete. It is designed to take biology apart, but not to put it back together. Reductionism produces vast numbers of isolated parts that have lost an essential property - connectivity. All complexities define themselves mathematically with parts and connections and biology is no exception. Remove the connections, and biology ceases to exist as a complexity. Consequently, our literature contains an enormous amount of information about parts, but surprisingly little about biology as it actually exists. Curiously, this basic truth remains largely unappreciated within the biology community.

Fortunately, the problem has a simple solution. We have the parts, but not the connections. Put the connections back, and we put back complexity. In return, we gain access to the ultimate problem solver – biology expressed as a mathematical pheno-

type. The following recommendations serve as helpful guidelines.

9.4 Recommendations

9.4.1 Theory Structure

- Begin to make the transition to a new theory structure for biology, one that combines mathematically the principles of both reductionism and complexity.
- Encourage the adoption of a complex approach to problem solving across all segments of the biology enterprise.

9.4.2 Technology

- Make the transition from small to big data.
- Encourage biology to become an information science, one driven by data instead of methods.
- Move published data onto a relational database platform or a suitable equivalent.
- Support the development of visual methods for analyzing large data sets.

9.4.3 Publication

- Adopt a two-tier system of publishing in the biological sciences – one for data and the other for traditional manuscripts.
- Improve the impact, value, and credibility of funded research by publishing data online.
- While biological databases clearly belong in the public domain, private companies will continue to serve an essential role in reviewing, editing, publishing, distributing, and archiving biological research.

9.4.4 Data Management

- Store data collected with reductionist methods in the tables of relational databases.
- Translate isolated data into ratios, mathematical markers, et cetera; store them in database tables.

- Use ratio based data sets – instead of concentrations and absolute value - to identify change and to characterize phenotypes.
- Use standardized templates when transforming and interpreting data sets.
- Provide specific guidelines for submitting data from the basic and clinical sciences to universal databases.
- Provide online facilities for entering, storing, and retrieving peer reviewed research data.
- Encourage a robust, quantitative approach to biological complexity by adding stereology to the biology curriculum.

9.4.5 Data Interpretation

- Interpret new research data within the framework of universal biology databases.
- Interpret data as connected patterns, rather than isolated data points.
- Interpret data within the framework of parallel complexities – or their equivalent.
- Encourage the reuse and reinterpretation of published data.

9.4.6 Support

- Develop comprehensive and long term funding for online literature databases.
- Use a common database design for the basic and clinical sciences.
- Introduce universal data compatibility into the design of information systems.
- Introduce something akin to Legacy Grants for senior investigators to move their published data into databases – as envisaged by the Biomatrix Project (Morowitz and Smith, 1987).
- Encourage the development of large-scale databases for biomedical research.
- Embark on a program designed to create an information space for the phenotype similar in size and importance to the one we already have for the genotype.

Epilogue

When we take a hard look at our current reductionist approach to biology, we discover that the data we collect are often mathematically unstable. The reason is simple. Biology runs mathematically as a complexity wherein both parts and connections are in play. In contrast, we remove the connections and the complexity and then try to run our business with just the parts. Consequently, our uneven record of accomplishment as a science should not come as a surprise. We have become very good at studying biological parts, but know surprisingly little about how to study biology as a complexity.

With the benefit of hindsight, we can risk a sweeping generalization. Biology – as a science – is more often than not incapable of delivering accurate and reproducible information about phenotypes – including the way they change. Our traditional methods are simply not up to the task. They can inject so much noise that the original biological data can become unintelligible. Biases contributed by faulty sampling, misinterpreted data, ignored valences, post-mortem distortions, and overblown biological variations only begin to tell the story. In fact, our methods create a vast complexity of their own, one that we have come to accept as a cost of doing business.

The cost, however, is far too high. Our self-imposed biases create a barrier, standing between biology and our charge to study it. It makes biology look disorganized and indecisive, when, in fact, it is a truly elegant system based on rules and running – as it must – on a mathematical engine. Remove – or at least minimize – this barrier and biology becomes a mathematical puzzle that we can learn to solve and replicate. Parallel complexities, which provide access to this puzzle, are already allowing us to harmonize published data, unravel the diagnosis-prediction problem, manage the crippling volume distortion issue of stereology, and explore the nature of the disease process in the human brain.

What did we learn by playing the complexity game with biology?

We learned how to make transitions – from static journals to dynamic databases, from a methods driven science to a data driven science, from subjective data to objective data, from disconnected data to connected data, from noisy data to quiet data, from distorted data to corrected data, from small data to big, from imagined simplicity to complex reality, and from chaos to order. With each transition, we honed our skills as problem solvers.

Why is it so important to approach biology as a complexity?

Biology knows how to solve – or at least manage successfully – a wide range of extremely difficult problems, many of which we cannot even begin to imagine. It accomplishes these remarkable feats within the framework of a rule-based system – hidden largely from view because of biology's fondness for wrapping its complexities within complexities. The simplest way of understanding such complexity is to become that complexity. Complexity theory, which allows us to test this idea with real world data, delivers a proxy for biology in the form of a parallel complexity. Starting with a catalogue of published data (IBVD), we can calculate ratios and then tune them deliberately to achieve specific objectives. Since the success of this approach depends on data being as close to the original as possible, living patients become our most reliable source. They give us a gold standard that we can defend both theoretically and empirically.

Perhaps, the most compelling argument for engaging complexity is that it allows us to learn the rules of the games we want to play. Knowing biology's rules provides access to large amounts of otherwise privileged information. Finding just a few rules has already provided enough information to reorganize the biology literature, minimize multiple sources of bias, and identify workable solutions to problems long considered as mission critical. The real prize, however, will be the patterns, algorithms and mathematics biology uses to manage and advance its relentless success as a complexity. Just imagine, for

example, a world in which we understood how to optimize nested complexities containing astronomical numbers of interacting parts and connections.

Given that our complexity game began with reductionism, it seems only fair to end with it as well. Reductionist theory has made a remarkable contribution to biology by allowing us to understand the many parts that make up living systems. Since we now know the names, values, and functions of the parts in a variety of settings, the theory has done its job. The time has come to move on and take the next step. This requires defining a new set of goals supported by new theory structures.

Since any new theory in biology must draw its strength from empirical data, it automatically becomes an extension of our current reductionist theory. Complexity becomes the logical choice because we can put the biology literature back in play by simply returning the connections to its mountains of isolated parts. Moreover, a complexity theory built upon the rules of an already successful biology finds itself in a position ready to deliver innovation throughout the enterprise. Lest we forget, the point in playing the game with biology is not just to win, but also to identify launching pads that can ensure our future success.

Our story is now well under way. The commitment made at the outset to get ourselves into as much trouble as possible provided exactly the incentive we needed to find out what was broken and how to fix it. We discovered that the same data viewed through different lenses delivers different outcomes. Complex problems viewed through simple lenses produce blurred images. View these same complex problems through complex lenses and the solutions snap into focus with the expected crispness of a mathematical science. Such is the argument put forward by this narrative. It maintains that biology becomes an objective science when it behaves like one. This includes having a theory structure with first principles firmly based on both theoretical and empirical arguments.

Even a casual reader scanning these chapters can see that reductionist theory represents a dumbing down approach to investigative biology. In retro-

spect, we seem to have invested so much for so long to learn so little. Although every competent knows that the only way to get smarter is to play a smarter opponent, we have stubbornly done quite the opposite. We repeatedly choose simplicity over complexity. By rejecting this conventional wisdom, we become free to chart new and more challenging directions. As our databases grow in size and scope, we can and should aspire to emulate biology in ways yet unimagined.

Biology becomes easier to understand when we treat its complexity as a mathematical jigsaw puzzle, with solutions in n -dimensional space. Starting with the one-dimensional strings of mathematical markers, we can reconstitute complexities at higher dimensions by simply snapping together the pieces. Such an approach allows our published data to form two general sets of patterns, one local and the other global. Although both sets can identify first principles, the global ones offer the advantage of reproducibility and predictability. Under the auspices of complexity theory, every publication becomes part of the larger whole and the biology enterprise grows smarter and more effective with each passing year. In effect, the theory structure defaults to success.

Biology also teaches us how to work faster and smarter. Under reductionist rules, it took more than three years to produce a stable database for the stereology literature (1998-2001) and even longer to extract informative patterns (2001-2011). By taking our rules from biology, however, it took less than a week to generate a universal database from the IBVD, three months to find a solution to the diagnosis problem, a week to parse the disorders of the brain, and a few days to come up with workable solutions to the volume distortion problems of stereology.

As scientists, we find ourselves in the curious business of having to invent the future. We do this by discovering new theory structures that allow us to solve otherwise intractable problems and to puncture the tired assumptions of old theories. Discovery, however, can become a very slippery business because it tends to trigger unintended consequences. Recall what happened as our story unfolded. In a very rudimentary way, we used published data to

create software devices that allowed us to set up a communication interface with biology. Such an outcome was possible because stereology allowed us to quantify all the parts that biology uses to build itself. First, however, we had to discover that complexity emerges only after our data reach a critical mass. By allowing large amounts of data to connect and interact, patterns appeared that began to explain how biology orders itself and manages its business. In turn, these patterns led to principles that gave us the all-important confidence to ask and answer harder questions. Instead of having to game the system as an outsider, we became an insider by adapting to the system we were trying to understand. By establishing mathematical links to biology, we end up on

the same page, share the same view of reality, and become allies. Biology's job is to assemble and manage a dazzling array of phenotypes, each one defined explicitly as a complexity of parts and connections. Our job becomes one of promoting the success of biology by sharing its problems and helping to find best solutions. This partnership, which identifies a core principle of complexity theory, will contribute importantly to the long-term success of biology and of our enterprise.

When we take biology apart, we see one thing, but when we put it back together, we see something entirely different.

Glossary

Working Definitions

ABSOLUTE DATA – Data expressed as a volume, surface, length, or number.

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.

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.

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. To wit, the turbulence created by a butterfly triggers a storm far away.

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.

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.

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.

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.

DATA PAIR – A ratio of two numerical values, which may include the names of the parts.

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.

DESIGN CODES – Include ratios formed by dividing experimental by control values. They identify patterns of change.

DISECTOR – A design-based method of stereology that uses an unbiased sampling frame to estimate the numerical density (N/V) of particles.

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.

ENTERPRISE BIOLOGY SOFTWARE PROJECT (EBSP) – A project designed specifically to speed learning and discovery in the life sciences.

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; a systematic random sampling method.

FRONT-END – User interface; the part of a software program with which the user interacts.

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.

IBVD – Internet Brain Volume Database

INTACT TISSUE – Undamaged; unaltered.

LADDER EQUATION – An exponential equation summarizing a set of rung (power) equations.

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 capture units of complexity specific to a given phenotypic state.

METHODS-DRIVEN – An activity compelled by methods.

NESTED COMPLEXITY – Complexity embedded in complexity. Unfolding and refolding nested com-

plexity 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.

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 a given paper graphically.

PARALLEL COMPLEXITY – A collection of mathematical markers serving as a proxy for biology; a proxy designed with a specific goal in mind (e.g., diagnosis).

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.

PLAYING FIELD – An database platform for playing complexity games with properties specified according to the game's rules; a field designed to solve a specific problem.

POWERBUILDER – An integrated development environment distributed by Sybase, Inc. (Emeryville, CA).

PROXY – A substitute for another.

QUADRUPLER MARKER – A mathematical marker consisting of four alpha and four numeric components; expressed as a numerical ratio.

QUERY BY EXAMPLE (QBE) – Query by example; a database query based on the items selected.

RATIO – Relative magnitudes of two or more quantities.

REDUCTIONIST THEORY – Assumes that complex systems can be completely understood in terms of their individual components (parts).

REGRESSION EQUATION – The relationship between values X and Y from which the most probable value of Y can be predicted from X.

REPETOIRE EQUATION – Defines the quantitative relationship of values X to Y, wherein both the slope and the R^2 of a power curve approach one.

REPRESENTATIVE SAMPLE – A population that accurately reflects the members of the entire population.

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.

SCIENCE – Extends knowledge of principles and causes.

STEREOLOGY – A collection of mathematical methods for estimating structures quantitatively.

STOICHIOMETRY – Relationships existing as a ratio of small integers.

SUBJECTIVE – Coming more from the observer than from observations.

THEORY – A well-substantiated explanation of some aspect of the natural world.

TRIPLET MARKER – A mathematical marker consisting of three alpha and three numeric components; expressed as a numerical ratio.

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.

UNIVERSAL BIOLOGY DATABASE – Contains biological data expressed as ratios; a unified data set derived from the biology literature.

VALENCE – An ability of a given part to connect to the same part in different ratios.

Bibliography

Abbott, E.A. 1991(New Material) Flatland: A Romance of Many Dimensions. Princeton University Press, Princeton, N.J. 104 pp.

Adami, C. 2015 The Information Theory of Life. Quanta Magazine, November 19, 2015. <https://www.quantamagazine.org/20151119-life-is-information-adami/>

Agostini A., Benuzzi F., Filippini N., Bertani A., Scarcelli A., Farinelli V., Marchetta C., Calabrese C., Rizzello F., Gionchetti P., Ercolani M., Campieri M. and P. Nichelli. New insights into the brain involvement in patients with Crohn's disease: a voxel-based morphometry study. *Neurogastroenterol Motil.* 2012 Sep 23. doi: 10.1111/nmo.12017.

Andersen, B.B. and B. Pakkenberg. 2003 Stereological quantitation in cerebella from people with schizophrenia. *Br J Psychiatry* 182: 354-361.

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.

Baker, K.G., A.J. Harding, G.M. Halliday, J.J. Krill, and C.G. Harper. 1999 Neuronal loss in functional zones of the cerebellum of chronic alcoholics with and without Wernicke's encephalopathy. *Neuroscience* 91: 429-438.

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. 2007A Rule Book: Guidelines to a Mathematical Biology. Edition 1.0, © 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. 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 J. M. Bluhm 1992 Database literature review: A new tool for experimental biology. *Mathl. Comput. Modelling*, 16:11-35.
- Bolender, R. P., Hyde, D. M., and R. T. DeHoff. 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): L521-L548.
- Borson S., Scanlan J., Friedman S., Zuhr E., Fields J., Aylward E., Mahurin R., Richards T., Anzai Y., Yukawa M., and S. Yeh. Modeling the impact of COPD on the brain. *Int J Chron Obstruct Pulmon Dis.* 2008 September; 3(3): 429–434.
- 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., and B.P. Lanphear. 2008 Decreased Brain Volume in Adults with Childhood Lead Exposure. *PLoS Med* 5(5): e112.
- Clarence L., Edwards S., Gong Q., Roberts N. and L.D. Blumhardt. Three dimensional MRI estimates of brain and spinal cord atrophy in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 1999;66:323-330 doi:10.1136/jnnp.66.3.323.
- 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>
- Committee on Models for Biomedical Research 1985 Models for Biomedical Research A new Perspective, National Research Council, National Academy Press, Washington, D.C.
- Cruz-Orive, L. M. and E. R. Weibel 1990 Recent stereological methods for cell biology: a brief survey. *Am. J. Physiol.* 258 (Lung Cell. Mol. Physiol. 2) L148-L156.
- Goldstein, J.M., Goodman, J.M., Seidman, L.J., Kennedy, D.N., Makris, N., Lee, H., Tourville, J., Caviness, F., Farcone, S.V. and M.T. Tsuang. 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., and T.T. Yang. 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, and M. J. West. 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.
- Hagmann P., Cammoun L., Gigandet X., Meuli R., Honey C.J., Van J. Wedeen, V.J., and O. Sporns. 2008 Mapping the Structural Core of Human Cerebral Cortex. *PLoS Biol* 6(7): 1479-1493. doi:10.1371/journal.pbio.0060159
- Harding A.J., Wong A., Svoboda M., Kril J.J., and G.M. Halliday. 1997 Chronic alcohol consumption does not cause hippocampal neuron loss in humans. *Hippocampus* 7:78-87.
- Harding, A.J., Lakay B., and G.M. Halliday. 2002 Selective hippocampal neuron loss in dementia with Lewy bodies. *Ann Neurol* 51: 125-8.
- Herting M.M., Gautam P., Spielberg J.M., Kan E, Dahl R.E., and E. R. Sowell. 2014 The role of testosterone and estradiol in brain volume changes across adolescence: A longitudinal structural MRI study. *Human Brain Mapping* Volume 35, Issue 11, 5633–5645.
- Ioannidis, J.P.A. 2005 Why most published research findings are false. *PLoS Med* 2: e124.
- Kauffman, S. At Home in the Universe. 1995 Oxford University Press, New York.

- Keller S.S. and N. Roberts. Measurement of brain volume using MRI: software, techniques, choices and prerequisites. *J Anthropol Sci.* 2009;87:127-51.
- Kennedy D.N., Hodge S.M., Gao Y., Frazier J.A., and C. Haselgrove. The internet brain volume database: a public resource for storage and retrieval of volumetric data. *Neuroinformatics.* 2012 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., and 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.
- Klein-Szanto, A.J. 1977 Stereologic baseline data of normal human epidermis. *J Invest Dermatol* 68: 73-78.
- 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 T. Smith 1987 Report of the Matrix of Biological Knowledge Workshop, Santa Fe, N.M., Santa Fe Institute.
- Nagai M., Hoshida S., and K. Kario. The insular cortex and cardiovascular system: a new insight into the brain-heart axis. *J Am Soc Hypertens.* 2010 Jul-Aug; 4(4):174-82.
- Nikicic H., Kasprzak A., and L.K. Malendowicz. 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.
- Pérez-Dueñas B., Pujol J., Soriano-Mas C., Ortiz H., Artuch R., Vilaseca M.A., Campistol J. Global and regional volume changes in the brains of patients with phenylketonuria. *Neurology*, Apr 2006; 66: 1074 - 1078.
- 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, and David N. Kennedy. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics.* 2012;6:9. doi:10.3389/fninf.2012.00009.
- Seecharan, D.J., Kulkarni, A.L., Lu, L., Rosen, G.D., and R.W. Williams. 2003 Genetic control of interconnected neuronal populations in the mouse primary visual system. *Neurosci* 23: 11178-88.
- Sterio D. C. 1984 The unbiased estimation of number and sizes of arbitrary particles using the disector. *J Microsc.* 134: 127-136.
- 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., and G.E. Alexander 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., and W.P. Mali 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.
- 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.
- West, M.J. 2012 Basic Stereology for Biologists and Neuroscientists, Cold Spring Harbor Laboratory Press, New York.

Index

- abnormal markers, 57, 82
- absolute values, 15, 23, 24, 25, 29, 31, 41, 42, 48, 64, 92
- absolute volume, 24
- Access, 68
- ADHD, 55, 58, 84, 85
- algorithms, 8, 10, 17, 27, 32, 98, 103
- Artificial Complexity, 63
- best practices, 64
- biological blueprint, 38, 39, 40, 41, 42
- biological complexity, 8, 9, 11, 13, 14, 15, 17, 28, 33, 35, 38, 64, 68, 97, 98, 101
- biological hierarchy, 16, 31, 32, 41, 52
- biological variation, 8, 9, 15, 17, 25, 29, 35, 41, 93, 94, 95, 96
- change, 10, 16, 23, 25, 26, 27, 30, 31, 33, 34, 35, 41, 42, 46, 51, 55, 58, 62, 65, 84, 92, 95, 100, 102, 103, 106, 107
- chaos theory, 15, 37, 106
- clouds of points, 55, 93
- CommunityGraphPlot, 80, 82
- complex patterns, 52, 53
- complexity, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 24, 25, 28, 29, 30, 31, 35, 37, 39, 45, 46, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 62, 63, 64, 65, 66, 67, 68, 80, 81, 82, 91, 92, 93, 94, 95, 97, 98, 99, 100, 101, 102, 103, 104, 107, 108
- complexity theory, 16, 49, 64, 67, 91, 94, 104
- concentrations, 10, 24, 25, 29, 31, 33, 41, 48, 92, 96, 102, 106
- correction factor, 65, 66, 67
- Counting Molecules, 24
- data cage, 76, 77, 78, 79
- data pairs, 15, 29, 30, 31, 32, 33, 36, 39, 40, 41, 43, 44, 45, 49, 53, 60, 61, 63, 69, 106
- data ratios, 14, 15, 44
- data triplets, 44
- decimal repertoire equations, 36, 37, 40
- decimal repertoire ratio, 53
- decimal repertoire ratios, 38, 40
- decimal steps, 36, 107
- descriptive science, 10, 12, 96
- design code equations, 26, 28, 42
- diagnosis, 37, 41, 43, 53, 54, 57, 60, 61, 65, 66, 68, 70, 71, 72, 73, 85, 91, 98, 103, 104, 108, 110
- disease process, 16, 39, 56, 57, 80, 81, 87, 90, 91, 98, 103
- disector method, 66
- disorders, 16, 43, 52, 53, 54, 55, 56, 57, 58, 59, 60, 64, 67, 69, 70, 71, 72, 73, 74, 75, 76, 77, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 98, 104, 107
- distorted volumes problem, 66
- distortions, 13, 14, 16, 25, 52, 62, 63, 64, 65, 66, 67, 92, 96, 98, 99, 103
- dominant central organizer, 47
- duplicates, 54, 61, 62, 63, 65, 81, 87, 90
- edge of chaos, 15, 94
- Enterprise Biology Software, 11, 16, 17, 101, 109, 110
- Excel, 31, 53, 68
- false positives, 68, 74
- Fibonacci, 18
- first principle, 29, 33, 38, 41, 47, 50, 67, 80, 107
- First Principles, 97
- fractionator, 64, 66, 93, 110
- generalizations, 11, 25, 26, 38, 40, 56, 59, 68, 101
- genotype, 39, 42
- global, 14, 15, 17, 34, 37, 44, 49, 51, 56, 58, 59, 61, 62, 80, 95, 98, 99
- Growth Kinetics, 38
- hierarchy equations, 24, 64, 92
- IBVD, 52, 53, 65, 70, 80, 90, 93, 94, 101, 103, 104, 107
- isogenic strains, 36, 37
- isolated data, 10, 15, 16, 17, 51, 93, 94, 102
- Ladder Equations, 32
- local, 14, 15, 37, 49, 51, 56, 68, 98, 99
- log growth, 38
- logical database model, 20
- Mathematica, 49, 53, 68, 80, 82
- Mathematical Mapping, 49
- mathematical markers, 16, 18, 52, 53, 54, 55, 56, 58, 60, 61, 62, 63, 64, 65, 68, 69, 80, 82, 91, 98, 102, 108
- methodological bias, 29

modules, 16, 18, 82, 83, 84, 91, 100
MRI, 11, 13, 16, 49, 53, 56, 59, 60, 61, 62, 63, 65, 94, 95, 101, 110, 111
optical volume fractionator, 66
organism codes, 45, 46, 49
paradox, 18
parallel complexity, 9, 15, 28, 52, 56, 59, 66, 68, 80, 81, 82, 91, 95, 97, 98, 103, 106
parallel regression curves, 33
Patterns, 17, 25, 36, 42, 51, 61, 62
phenotype, 8, 9, 12, 14, 17, 27, 32, 35, 38, 39, 41, 42, 49, 52, 53, 56, 69, 80, 91, 95, 101, 102
playing field, 13, 14, 28, 29, 49, 59
polynomial, 41, 42, 43, 52, 106
post-mortem, 13, 14, 16, 35, 39, 43, 49, 52, 60, 61, 62, 63, 64, 65, 66, 67, 68, 92, 93, 96, 98, 99, 101, 103
PowerBuilder, 68
Prime Movers, 87
proxy, 15, 103, 108
quadruplet marker, 68
quadruplet markers, 70, 71, 73, 74, 75, 76, 81
quantitative science, 9, 10, 101
query by example, 40
ratio data, 32, 36, 38, 44
Recommendations, 101, 102
reductionism, 8, 12, 13, 16, 17, 24, 25, 27, 31, 51, 68, 94, 95, 102, 104
reductionist theory, 10, 27, 28, 29, 60, 104
relational database, 11, 14, 21, 27, 28, 53, 102
Repertoire equations, 31
Reverse Engineering, 35
rules, 8, 9, 10, 12, 13, 14, 16, 17, 20, 25, 26, 29, 30, 38, 39, 41, 45, 49, 59, 80, 92, 95, 97, 98, 101, 103, 104, 106, 107, 108
Rung Equations, 33
scatter plots, 55, 93
shared markers, 69, 70, 72, 80, 83, 84, 91
shrinkage, 65
simplicity, 8, 12, 18, 92, 101, 103, 104
simulators, 33, 34, 35
small whole numbers, 41
standards, 17, 53, 65, 67, 70
stereology, 11, 14, 15, 16, 24, 25, 28, 36, 39, 60, 62, 63, 64, 65, 66, 93, 97, 101, 103, 104, 107
Stereology Literature Database, 60
stoichiometry, 39, 45, 51, 98
Structured Query Language, 40
swelling, 65
symptoms, 43, 53, 80, 85, 86, 87, 91
Technology Shift, 68
Theory of Biological Complexity, 98
Theory Structure, 98, 102
triangular patterns, 50
triplets, 44, 45, 46, 49, 52, 53, 69
two complexities, 13, 62
unbiased, 14, 25, 66, 97, 98, 107, 111
unfolding, 11, 32, 38, 50, 56, 59
unique markers, 72, 74, 76, 78
universal biology database, 15, 29, 30, 32, 38, 39, 101
valences, 15, 18, 31, 33, 34, 35, 37, 39, 96, 100, 103
volume dependent, 66, 67, 93
volume distortions, 25
volume independent, 64, 66, 67, 93