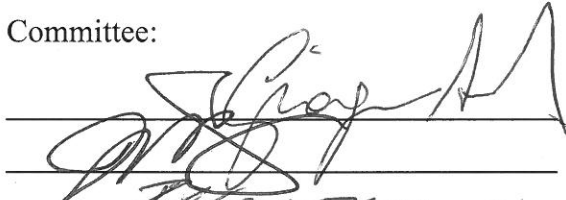


COMPARATIVE TOPOLOGICAL ANALYSIS OF NEURONAL ARBORS VIA
SEQUENCE REPRESENTATION AND ALIGNMENT

by

Todd Aaron Gillette
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Neuroscience

Committee:



Dr. Giorgio A. Ascoli, Dissertation Director



Dr. James Olds, Committee Member



Dr. Iosif Vaisman, Committee Member



Dr. Kenneth De Jong, Director, Krasnow
Institute for Advanced Study



Dr. Donna Fox, Associate Dean for
Academic and Student Affairs, College of
Science



Dr. Peggy Agouris, Dean, College of
Science

Date: 4/22/15

Spring Semester 2015
George Mason University
Fairfax, VA

Comparative Topological Analysis of Neuronal Arbors via Sequence Representation and
Alignment

A Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

Todd Aaron Gillette
Bachelor of Science
Swarthmore College, 2003

Director: Giorgio A. Ascoli, Professor
Department of Neuroscience

Spring Semester 2015
George Mason University
Fairfax, VA



This work is licensed under a [creative commons attribution-noncommercial 3.0 unported license](https://creativecommons.org/licenses/by-nc/3.0/).

DEDICATION

This is dedicated to my parents, Cheryl and Herb Gillette, who taught me to value learning, exploration, and determination; to work through challenges with “eyes on the prize” and never give up. You supported me, pushed me to do take the challenging but rewarding path, and always expected me to succeed.

ACKNOWLEDGEMENTS

I would like to thank Giorgio Ascoli for providing guidance and diverse opportunities, and for driving me to sharpen my organization and communication skills. Thanks to my other committee members for their unique contributions: Jim Olds for his support throughout my time at Krasnow and his wise words and experience regarding science and academia; and Iosif Vaisman for providing valuable insight at a vital stage of my work.. Other former committee members who had to depart my committee also deserve my thanks: Ted Dumas for teaching me the messy side of science and providing the advice to not lose track of the biology while working on methods and data analysis; John Grefenstette for engaging early on with the strange idea of encoding neurons as sequences, and offering direction in my efforts to learn informatics techniques without classroom experience.

Thank you to all of the Center for Neural Informatics, Structures, and Plasticity (CN3), past and present, most notably Kerry Brown for being an excellent partner on two projects, Parsa Hosseini for key bioinformatics context and input, and long-time supportive friends and colleagues in the CN3 and Neuroscience program: Rebekah Evans, Michele Ferrante, Ruchi Parekh, Sarah Clark, Sriraman Damodaran, Duncan Donohue, Sarah Hawes, Jacob Nordman, Sridevi Polavaram, Christopher Rees, Erin Sanders, Susan Wright, among many others. CN3 PI, neuroscience program director, and professor of several of my favorite classes, Avrama Blackwell also deserves special mention.

I would also like to recognize the Krasnow Institute administrative staff, all of whom have provided incredible support, particularly Jennifer Sturgis, Jane Wendelin, Joey Carls, Carolyn Payne, and Stu Mackenzie; as has the staff at College of Science, notably Melissa Hayes.

Last but certainly not least, I acknowledge and have been blessed by the support and patience of my incredible fiancée, Laura Drachsler, who over the years has put up with my many late nights at work, disorganized descriptions of whatever challenges I might have been facing any given week, and erroneous predictions regarding my year of graduation.

TABLE OF CONTENTS

	Page
List of Tables	ix
List of Figures	x
List of Abbreviations	xii
Abstract	xiii
Chapter One: Introduction	1
Chapter Two: Measuring and modeling morphology: how dendrites take shape	6
Abstract	6
Introduction	7
Morphological Data and Computational Capabilities	10
Quantitative Measures and Functional Implications	14
Size and Geometry	16
Topology and Complexity	21
Real Morphologies and their Properties	31
Molecular Underpinnings of Dendritic Development	35
Cytoskeleton	35
Regulatory Influences and Signaling Molecules	38
Biologically Based Models	41
Models of Growth and Morphology	44
Determining the Rules of Growth	45
Constraints and Functions	49
Statistical Models	57
Dendritic Morphology into the Future	63
Acknowledgements	65
Chapter Three: Towards High-Throughput Data Processing by Promoting Development of Automated Morphological Reconstruction	66

3.1. DIADEMchallenge.Org: A compendium of resources fostering the continuous development of automated neuronal reconstruction	66
3.2. The DIADEM Metric: Comparing Multiple Reconstructions of the Same Neuron	68
Abstract.....	68
Introduction	69
Previous Relevant Work.....	74
Description of the DIADEM Metric.....	77
Human Expert Qualitative Scores and Interaction Time.....	84
Utility of the DIADEM Metric	87
Discussion.....	93
Acknowledgements	97
Chapter Four: Topological characterization of neuronal arbor morphology via sequence representation. I. Motif analysis.....	98
Abstract	Error! Bookmark not defined.
Introduction	99
Methods	101
Neurite trees to sequences	101
K-mers	104
Constrained tree-sequence generation	107
Tree growth methods.....	107
The NeuroMorpho.Org dataset.....	108
Results	111
Length and node-type proportions across arbors types	111
Motifs and Anti-motifs	116
Discussion	122
Data biases	124
Alternative encodings and traversals	125
Model validation and additional applications.....	126
Conclusions	127
Acknowledgements	127
Chapter Five: Topological characterization of neuronal arbor morphology via sequence representation. II. Global Alignment	128

Abstract	128
Introduction	129
Methods.....	132
Global tree-sequence alignment	135
Spatial embedding and classification	137
Multiple sequence alignment.....	139
NeuroMorpho.Org dataset	140
Results	140
Topology alignment space.....	141
Classification and comparison to morphometrics	143
Dendrite and axon clusters	145
Multiple sequence alignment and consensuses	149
Consensuses and motifs.....	153
Discussion	155
Biological and functional interpretations	155
Sequence alignment and expanded representations.....	157
Optimizing algorithms and analysis	160
Conclusion.....	162
Acknowledgements	163
Appendix One: Details of Global and Multiple alignment processes.....	164
Global sequence alignment	164
Score baseline and distance normalization.....	167
Additional methodological details	168
Non-metric MDS	168
Model-based clustering.....	169
Classification	170
Appendix Two: Quantifying neuronal size: Summing up trees and splitting the branch difference	172
Abstract	172
Introduction	172
Neuromorphology.....	173
Neuron size and shape.....	174
Cell growth in development	177

Morphological plasticity	178
Extrinsic vs. intrinsic influences.....	179
Qualitative and quantitative neuromorphological features	180
Beyond size: complexity and asymmetry.....	183
Partition asymmetry.....	184
Asymmetry alternatives	185
Asymmetry's complex relationships	188
Caulescence	189
Conclusion.....	195
Acknowledgements	195
Appendix Three: On Comparing Neuronal Morphologies with the Constrained Tree-edit-distance	196
Abstract	196
Body	197
Appendix Four: Statistical analysis and data mining of digital reconstructions of dendritic morphologies.....	204
Abstract	204
Introduction	205
Methods	210
Selection of Datasets and Morphometric Features for Analysis	210
Principal Component Analysis (PCA) and Cluster Analysis	213
Results	216
Variability of Dendritic Morphology and Comparison by Metadata	216
Extracting Primary Morphological Features by PCA and Cluster Models	220
Statistical Associations between Clusters and Metadata Combinations	224
Pairwise Morphometric Comparisons of Neuron Groups Identified by Cluster Analysis	230
Critical Assessment of Potential Confounds	234
Discussion	238
Acknowledgments	241
References	243
Biography.....	259

LIST OF TABLES

Table	Page
Table 3.1. Scores across data sets.	89
Table 3.2. Summary of correlations.	90
Table A4.1. Coefficients of variation of all L-Measure derived morphometric features.	213
Table A4.2. Morphometric loadings.	221
Table A4.3. Cluster metadata associations.	225
Table A4.4. Composition of the six morphological clusters in terms of their over- represented metadata groups.	229

LIST OF FIGURES

Figure	Page
Figure 2.1. Diversity in neuronal morphology.....	8
Figure 2.2. Morphological reconstruction.	12
Figure 2.3. Size and geometry features.....	16
Figure 2.4. Morphological effects on signal propagation and firing rates.....	18
Figure 2.5. Unlabeled binary tree shapes.....	23
Figure 2.6. Asymmetry.	24
Figure 2.7. Caulescence.	28
Figure 2.8. Sholl analysis design and plot.	30
Figure 2.9. Morphology and electrophysiology.....	34
Figure 2.10. Developmental molecules and mechanisms of dendritic growth and branching.....	36
Figure 2.11. Example of stimulus induced growth.	41
Figure 2.12. Modeling molecular-driven growth.....	43
Figure 2.13. Terminal and segmental growth.	46
Figure 2.14. Modeling extrinsic factors.	50
Figure 2.15. Functionally constrained modeling.	52
Figure 2.16. Divergence and convergence.....	56
Figure 2.17. Modeling somatopulsion (a).....	62
Figure 3.1. Effects of morphological features.	72
Figure 3.2. Tree registration and tree edit distance (TED).	76
Figure 3.3. The DIADEM metric.....	80
Figure 3.4. Matching paths through unregistered nodes.....	82
Figure 3.5. Determining excess nodes.	84
Figure 3.6. Neuromuscular projection axon rosette.....	84
Figure 3.7. Scoring method comparison.	88
Figure 3.8. Relationships between scoring methods.....	91
Figure 3.9. Tracing errors.	93
Figure 4.1. Converting morphologies to sequences.....	103
Figure 4.2. Measuring k -mers.	106
Figure 4.3. Tree size and complexity.....	110
Figure 4.4. Neurite size and node type features.....	114
Figure 4.5. Dimer analysis reflects terminal growth effects.....	119
Figure 4.6. Distinguishing trimers of axons and dendrites.....	121
Figure 4.7. Dimer motifs and anti-motifs by arbor type.....	121
Figure 4.8. Axon and dendrite trimer motifs and anti-motifs.....	121

Figure 5.1. Pairwise global sequence alignment.....	134
Figure 5.2. Alignment space and clustering.....	143
Figure 5.3. Dendrite clusters.....	146
Figure 5.4. Axon clusters.....	148
Figure 5.5. Multiple sequence alignment and consensuses.	152
Figure 5.6. Global consensus and local motifs.	154
Figure A1.1. Dynamic programming alignment of tree node types.	166
Figure A1.2. Normalizing the per-character alignment score.....	168
Figure A1.3. Multiple sequence alignment algorithm.	171
Figure A2.1. Diversity in neuronal size.	176
Figure A2.2. Binary tree properties.	185
Figure A2.3. Main paths in example neurons from NeuroMorpho.Org.	191
Figure A2.4. Global asymmetry versus caulescence.	194
Figure A3.1. Algorithm run-time.....	201
Figure A3.2. Run-time of queries by size.	202
Figure A4.1. Sample of NeuroMorpho.Org reconstructions representing the anatomical diversity of dendritic and axonal trees.	206
Figure A4.2: Search and download features available in NeuroMorpho.Org.	208
Figure A4.3. NeuroMorpho.Org v5.6 data is categorized along three major metadata dimensions, namely species, brain regions, and cell types.	217
Figure A4.4. Inter-group differences of representative morphometric features within each main metadata dimension.	219
Figure A4.5. Unsupervised cluster analysis of dendritic morphology.....	223
Figure A4.6. Similarities and differences of rodent retinal ganglion cells with other neurons within and across clusters.....	231
Figure A4.7. Rodent and primate cortical pyramidal cells show a distinct linear relationship between PC1 and PC3.....	234

LIST OF ABBREVIATIONS

Bayesian Information Criterion	BIC
Expectation Maximization	EM
Linear discriminant analysis	LDA
Long then Short tree traversal	LtS
Principal component analysis	PCA
Principal component	PC
Short then Long tree traversal	StL

ABSTRACT

COMPARATIVE TOPOLOGICAL ANALYSIS OF NEURONAL ARBORS VIA SEQUENCE REPRESENTATION AND ALIGNMENT

Todd Aaron Gillette, Ph.D.

George Mason University, 2015

Dissertation Director: Dr. Giorgio A. Ascoli

Neuronal morphology is a key mediator of neuronal function, defining the profile of connectivity and shaping signal integration and propagation. Reconstructing neurite processes is technically challenging and thus data has historically been relatively sparse. Data collection and curation along with more efficient and reliable data production methods provide opportunities for the application of informatics to find new relationships and more effectively explore the field. This dissertation presents a method for aiding the development of data production as well as a novel representation and set of analyses for extracting morphological patterns.

The DIADEM Challenge was organized for the purposes of determining the state of the art in automated neuronal reconstruction and what existing challenges remained. As one of the co-organizers of the Challenge, I developed the DIADEM metric, a tool designed

to measure the effectiveness of automated reconstruction algorithms by comparing resulting reconstructions to expert-produced gold standards and identifying errors of various types. It has been used in the DIADEM Challenge and in the testing of several algorithms since.

Further, this dissertation describes a topological sequence representation of neuronal trees amenable to various forms of sequence analysis, notably motif analysis, global pairwise alignment, clustering, and multiple sequence alignment. Motif analysis of neuronal arbors shows a large difference in bifurcation type proportions between axons and dendrites, but that relatively simple growth mechanisms account for most higher order motifs. Pairwise global alignment of topological sequences, modified from traditional sequence alignment to preserve tree relationships, enabled cluster analysis which displayed strong correspondence with known cell classes by cell type, species, and brain region. Multiple alignment of sequences in selected clusters enabled the extraction of conserved features, revealing mouse neocortical pyramidal cell axons and rodent neocortical dendritic-targeting interneurons to be substantially more asymmetric than perisomatic-targeting interneurons. With optimization techniques adapted from the field of genomic alignment, these methods compose a framework with the potential to be made orders of magnitude more efficient. Moreover, the framework is capable of handling expanded sequence representations that include additional branch features, enabling analysis of correspondence and joint conservation of various morphological characteristics.

CHAPTER ONE: INTRODUCTION

As the field of neuroscience progresses, new questions arise along with new types of experiments, data collection methods, formats, and analysis. Neuroinformatics approaches can be brought to bear, directly and indirectly advancing each component of research. A variety of research questions involve neuronal morphology as it provides clues to development and impacts function in terms of both dendritic computation and network connectivity. The compilation and curation of morphological data from a wide variety of species, cell types, brain regions, and experimental conditions has begun to allow reuse of data and exploration of relationships that would previously have required extensive effort and collaboration (Teeter and Stevens, 2011; Cuntz *et al.*, 2012; Polavaram *et al.*, 2014). Additionally, new imaging and data production methods (Chiang *et al.*, 2011; Jenett *et al.*, 2012) promise orders of magnitude more data akin to the era of big data experienced in genomics. This dissertation presents efforts towards enhancing data production methods along with novel representations and methods capable of taking advantage of larger and more diverse datasets to analyze neuronal classes for morphological patterns.

To provide context for the methods and analyses, Chapter Two: contains a background on neuronal morphology, focused primarily on dendrites. Morphological features, metrics used to capture them, and their relationships with electrophysiology are

discussed initially, followed by an exploration of the molecular mechanisms underlying development and plasticity. Further, a variety of phenomenologically and statistically based models are discussed which have been used for elucidating the various relationships between morphology, growth, and function.

The development of tools for producing data takes up an entire area of research. Chapter Three: describes the DIADEM Challenge, an effort to aid in the development of automated reconstruction tools. Labeling, imaging, and extracting features of interest in neuroanatomy began with the Golgi stain, basic light microscopy, and hard-drawn images. Camera lucida was an important advance that simplified the process by allowing a scientist to trace over a projection of a microscope image. Computer technologies have made storage and computer-assisted reconstruction possible, while the Allen Brain Institute brought the processes of tissue preparation and imaging into a high-throughput era in producing the Allen Brain Atlas. Similar approaches are being developed for imaging *Drosophila* brains at the single neuron level, and major projects are being discussed and developed in order to allow such high-throughput and high-resolution in the rodent and possibly human brains. Of the myriad challenges faced in such efforts, the reconstruction process is a major bottleneck.

The DIADEM Challenge, described in Chapter 3.1, was designed to gauge the state of the art in automated reconstruction and determine what issues remained to be overcome. As an organizer of the Challenge, I provided technical input as well as ideas and feedback on both the qualifier and finalist phases, including the competition conference. During the competition finals, I co-led a team of volunteers to track human

interface time between each team and their algorithms given the goal of high-quality automated reconstructions with minimal human intervention. Additionally, I led the development of the DIADEM metric (described in Chapter 3.2), one of the primary methods used for measuring the quality of automated reconstructions relative to gold standards in qualifier and final phases of the Challenge. The DIADEM metric is presented and compared to expert judgment. Since its use in the DIADEM Challenge, it has been used to aid development of and test the capabilities of new automated reconstruction algorithms.

Chapter Four: approaches the issue of data representation and forms of analysis with a novel sequence representation of morphology developed to allow for more sensitive analysis of topology and of combinations of morphological features. The representation, inspired by genomics, is amenable to various sequence analysis techniques. As proof of concept, and to explore the topological relationships between arbor and neurite types, a purely topological representation is used in a motif analysis. The analysis found substantial differences in node type proportions across arbor types, but much smaller differences between axons and dendrites in terms of relative prevalence of short sequences. Apical dendrites, on the other hand, are shown to differ substantially from both groups at each level of analysis, suggesting a strong topological relationship with spatial coverage and synaptic integration.

The local structural findings of motif analysis are followed by an exploration of global topological features achieved with global sequence alignment in Chapter Five:. Global alignment enables a detailed level of comparison between two neurites as opposed

to choosing a metric and taking its average across the branches. With a matrix of alignment scores as a measure of similarity, dominant features emerge upon embedding the neurites in an abstract multidimensional space. Clusters in that space indicate neurites with some functional relationship that can be investigated by determining the associations of those clusters with known metadata and by extracting the conserved features responsible for the clustering via multiple sequence alignment. Further details of the global and multiple sequence alignment processes are provided in Appendix One:.

The features of global alignment and local motifs prove to be complementary, though a variety of other related analyses are also made possible with a sequence representation. Particularly with more expansive encodings, searches for similarity can be improved and limited to alignment of local regions rather than only the entire neurite. Chapter Four: and Chapter Five: thus provide useful analysis of the relationships between topology, growth, and function, while providing new tools for further exploration of morphology in the coming era of big neuroscience data.

Three additional appendices are included, each a paper to which I contributed and related to neuronal morphology. The first appendix begins with a review of the field and concludes with the presentation of caulescence, a metric related to asymmetry that determines a main path and indicates its prominence. Appendix Two is a commentary I wrote that discusses the tree edit distance and its uses in neuroscience and beyond. The tree edit distance is a measure used in several fields involving tree structures, and one related to the global alignment method discussed in Chapter 5. The final appendix is a research article to which I contributed that presents an unsupervised cluster analysis of

neuronal dendrites based on morphometrics. The cluster and association analysis is similar to that used in Chapter Five. The appendices compliment the main text of the dissertation, providing greater depth into methods and analyses related to the main studies.

CHAPTER TWO: MEASURING AND MODELING MORPHOLOGY: HOW DENDRITES TAKE SHAPE

[With kind permission from Springer Science+Business Media: Computational Systems Neurobiology, Chapter 13, Measuring and modeling morphology: How dendrites take shape, Todd A. Gillette, Giorgio A. Ascoli, ed. Nicolas Le Novère, © Springer Science+Business Media Dordrecht 2012]

Abstract

Neuronal processes grow under a variety of constraints, both immediate and evolutionary. Their pattern of growth provides insight into their function. This chapter begins by reviewing morphological metrics used in analyses and computational models. Molecular mechanisms underlying growth and plasticity are then discussed, followed by several types of modeling approaches. Computer simulation of morphology can be used to describe and reproduce the statistics of neuronal types or to evaluate growth and functional hypotheses. For instance, models in which branching is probabilistically determined by diameter produce realistic virtual dendrites of most neuronal types, though more complicated statistical models are required for other types. Virtual dendrites grown under environmental and/or functional constraints are also discussed, offering a broad perspective on dendritic morphology.

Introduction

It is generally understood that for anything that is the result of an evolutionary process, form follows function. A streamlined body in both nature and the world of man-made machines usually indicates an object's function is to move quickly. Even such a seemingly mundane object such as a foot can say a great deal about the foot's owner. Water fowl feet are webbed for paddling, birds of prey have sharp talons for grabbing and killing their prey, many primates have feet with thumbs for traversing trees, and humans have particularly large feet to keep a more upright body well balanced. So, in order to understand function and the developmental process of dendrites, it is highly worthwhile to study their morphological attributes.

A look through the online neuronal reconstruction database NeuroMorpho.Org (Ascoli *et al.*, 2007) will provide one with a clear view of the great variety of dendritic and axonal morphologies. We see differences in size and shape from the planar and densely bifurcating dendrites of the cerebellar Purkinje cell (**Figure 2.1a**) to the polarized and conical shape of the pyramidal cell's basal and apical dendrites (**Figure 2.1b**). The cortical stellate cell, among many others, exhibits radial growth in all directions (**Figure 2.1c**). In the many visually distinctive neuronal classes one can easily see the diversity in neuron size, density, proportion, and complexity. Properties such as taper rates and asymmetry are harder to discern or compare by eye, but are equally important.

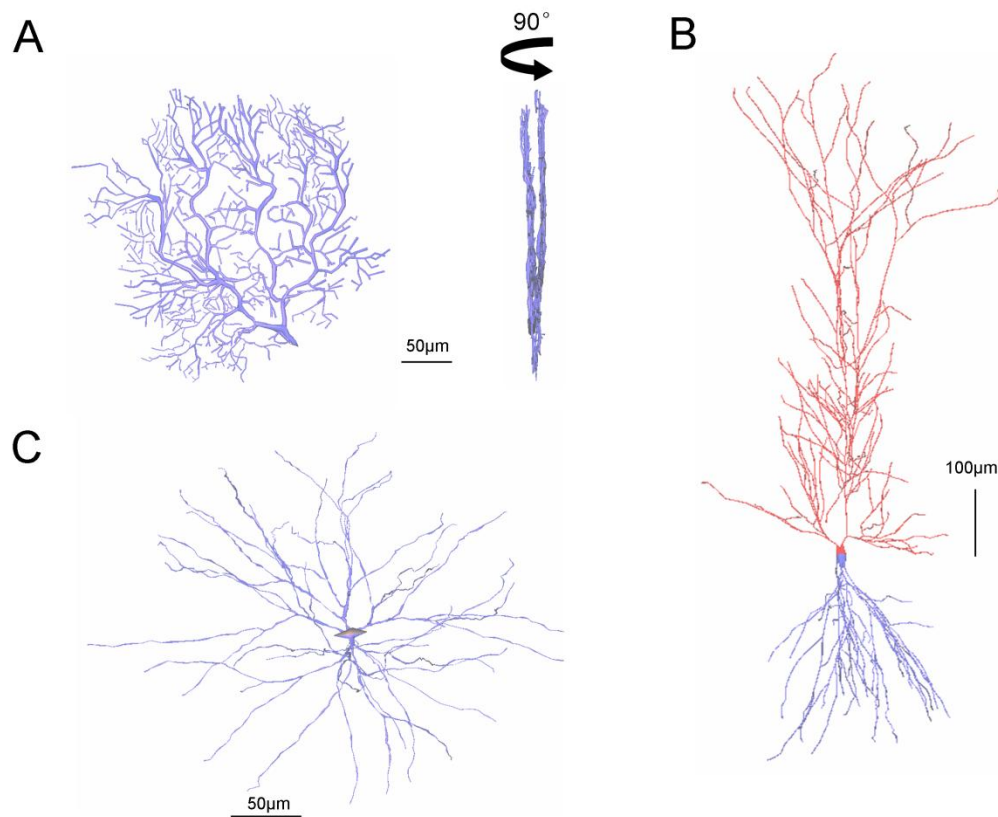


Figure 2.1. Diversity in neuronal morphology. **a.** Cerebellar Purkinje cell viewed with its largest area showing (*left*) and edge-on (*right*), exhibiting dense arborization and a planar shape (Rapp *et al.*, 1994) **b.** Hippocampal CA1 pyramidal cell with polarized basal (*blue*) and apical (*red*) dendrites (Ishizuka *et al.*, 1995). Growth is primarily within a conical region, particularly for the basal dendrites. **c.** Cortical stellate cell with radial dendritic arborization in all directions (Vetter *et al.*, 2001). Reconstructions retrieved via NeuroMorpho.Org (Ascoli *et al.*, 2007).

The effort to understand the complexities of dendrites may be largely attributed to the variety of functions they must perform. Different dendritic classes and even individual dendrites are characterized by specific morphological and functional properties. Aside from the particulars of how a dendrite must integrate its synaptic inputs, it must minimize metabolic costs while reaching existing and potential synaptic targets, make appropriate connections to form the network of which it is a part, and be capable of

plasticity required for various forms of learning. The morphology of a dendrite has evolved to accomplish these goals, but it may be that in the interest of accomplishing one goal, the morphology is detrimental towards other goals. In such cases, other properties may make up for the counterproductive effects.

For instance, while bifurcations are known to reduce the probability of the continuation of a spike (Sieck and Prakash, 1995; Spruston *et al.*, 1995; Williams and Stuart, 2000), it is not as clear whether this is useful or detrimental to signal processing. Both channel properties (e.g. density and conductance) and the ratio of parent to daughter segment diameter can impact failure probability (Migliore, 1996). If a discovery is made that a greater density of sodium channels or local enlargements of daughter branch diameter are common at bifurcations, then one could argue that neurons have developed mechanisms to reduce the probability of bifurcation spike failure as the increased channel density and daughter branch taper decrease the likelihood of failure. An even more interesting discovery would be of a mechanism providing control over local channel properties or dendritic diameter and therefore propagation failure probability. In subsection **Size and Geometry** the relationship between spike propagation and morphology will be further explored. By understanding how particular morphological features affect electrophysiology and how connectivity patterns affect morphology, we can better understand the function of various neuron types as well as how the overall morphology of a neuron and its non-morphological features contribute to function.

A wide variety of techniques are used to explore the relationships between morphology, development, and function. Electrophysiological experiments and models,

as well as molecular and other biophysical research, provide functional data. Imaging and reconstruction techniques provide a means to produce rich morphological data. Models fall into and across multiple categories, each with particular strengths for both testing and generating hypotheses of development and function. Morphometrics enable measurement of morphological features which in turn provides the capability to compare cell classes quantitatively and find correlations between morphology and function. Comparisons between cell classes can be extended to morphologies produced by models, providing a mechanism for their validation. This chapter will cover the tools and techniques used to study dendritic morphology, as well as many of the known and proposed underlying mechanisms responsible for growth and the associated functions achieved by dendrites.

Morphological Data and Computational Capabilities

Before delving into the deeper complexities of morphological models and experiments, it will be useful to understand how the source data is derived and organized. Computing power and algorithmic advances have provided phenomenal opportunities over the last decade for three dimensional digital reconstruction, complex analysis, and simulation of both dendritic morphology and electrophysiology. Applying information technology to the field is also having a great impact, with databases of morphology, physiology, and simulation models such as NeuroMorpho.Org (Ascoli *et al.*, 2007), NeuronDB (Mirsky *et al.*, 1998), and ModelDB (Hines *et al.*, 2004) respectively.

Three dimensional neuronal reconstructions are digitized representations of the morphological structure of neurons. These representations are made up of tapering cylinders (technically cylindrical frustums) that are connected in a tree structure to

represent a neuron's soma, dendrites, and/or axon. Some reconstructions may also include dendritic spine data, histological or anatomical boundaries, or other supplementary information. Currently, these reconstructions are created primarily by a human operator working either on a computer equipped microscope or on image stacks produced through a microscope for “offline” tracing (**Figure 2.2**). Semi-automated programs are available to assist human users, simplifying the process. These include the freely available Neuromantic (Neuromantic: the freeware neuronal reconstruction tool) the popular commercial Neurolucida (Glaser and Glaser, 1990); (MBF bioscience: neurolucida - neuron reconstruction) and Bitplane's FilamentTracer (Bitplane: neuron reconstruction - automatic neuron tracing, spine detection and analysis in 3D/4D). Fully automated programs exist as well, but are not yet widely employed as they have not been proven reliable on a representative variety of cell types, staining techniques, and microscopy. A competition, the Digital Reconstruction of Axonal and Dendritic Morphology (DIADEM) Challenge, was launched in April 2009 in order to promote development and test the capabilities of such programs (Gillette *et al.*, 2011a). Automated reconstruction may yield within a few years an explosion of morphological data and connectomic data. Each type of data is useful in its own right, but given the relevance of connectivity to morphology, both types of data will complement each other in both areas of research.

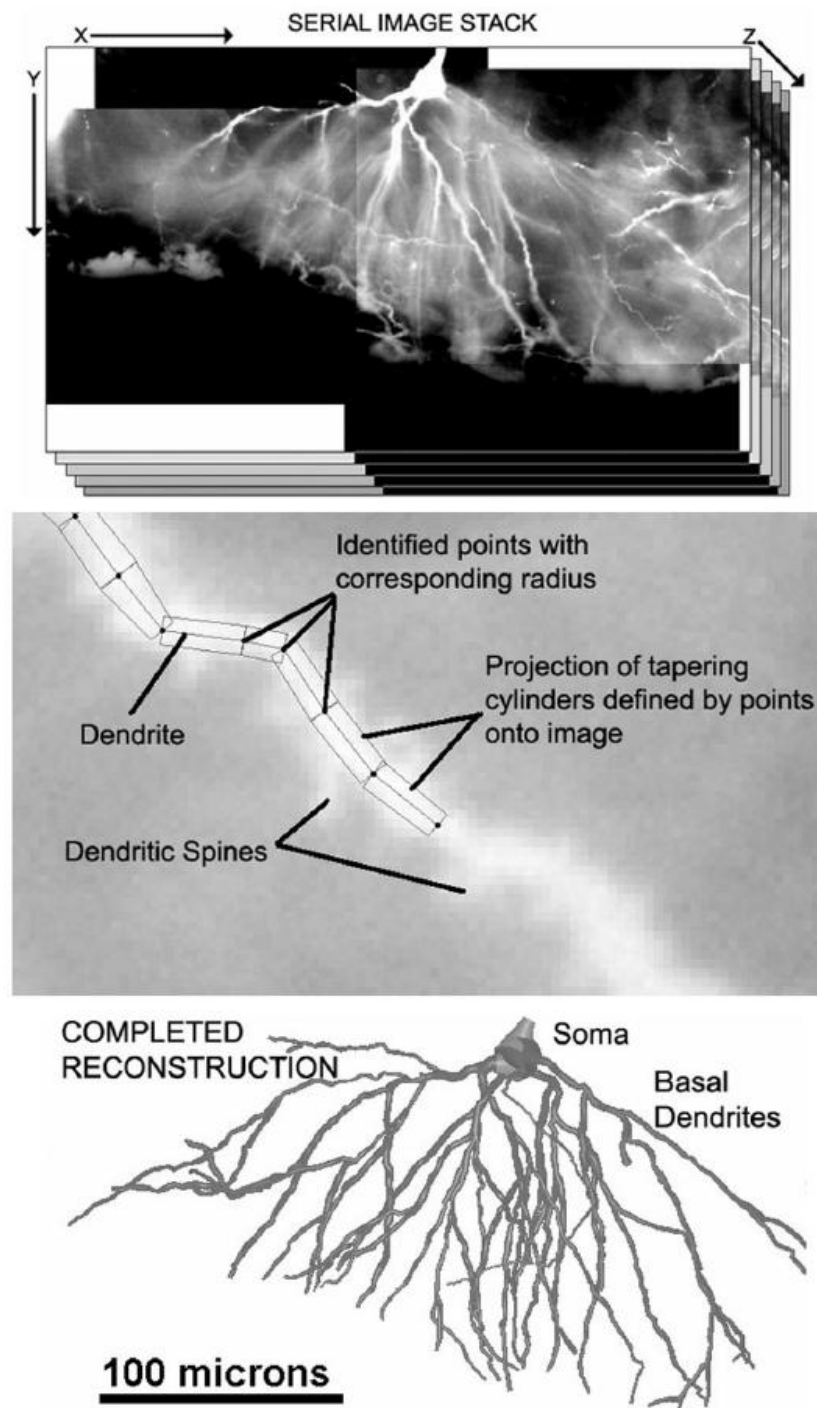


Figure 2.2. Morphological reconstruction. **Top.** Reconstructing a neuron begins with a three-dimensional image stack captured from a microscope. **Middle.** Compartments are reconstructed one by one along each neurite in an image. **Bottom.** Once completed, the full 3D reconstruction can be viewed and analyzed. (Adapted from with permission from [Myatt *et al.*, 2012]).

Open access databases, in combination with analytical tools, provide the opportunity for both broad canvases of morphologies and physiological data as well as more targeted analyses using data already acquired for other research. A variety of neuroscience related databases can be found at the Neuroscience Information Framework (Gardner *et al.*, 2008, <http://neurogateway.org/>). By sharing and making the appropriate data easy to find, the impact of the original research can be multiplied, both scientifically and in terms of citations (Ascoli, 2006). Prior to such databases, sharing data was possible but required more time and effort. Now data is available from multiple labs, and as the databases become better known and used, it will become more likely for a researcher to volunteer their data. In the case of NeuroMorpho.Org, this will create the prospect of more up-to-date and complete databases with morphologies across the spectrum of species, brain region, cell class, and even experimental conditions such as knock out, age, and behavioral experience. Analytical tools that can be coupled to such databases can multiply their impact. L-Measure is one such example; a freely available cross-platform tool with the ability to calculate and statistically analyze a broad array of morphometrics for any selected set of morphologies (Scorcioni *et al.*, 2008, <http://krasnow1.gmu.edu/cn3/>).

In addition to molecular and biological experimentation, computer simulation has been another vital tool for morphology research. Aside from electrophysiological models that make use of real dendritic morphologies, growth models in concert with morphological analysis techniques have allowed scientists to test developmental and functional hypotheses. There are many methods of modeling dendritic morphology, each

offering different benefits and each capable of addressing different questions. Some models use a particular mechanism of growth, such as bifurcation at terminal tips or growth due to external environmental factors. Other models use statistics to characterize the relationships between initial and emergent properties during the growth process. For example, one such model determines bifurcation and termination probability based on a segment's diameter. In such models “growth” does not necessarily imply developmental growth, but rather the process the model undergoes to generate a dendrite. These models are useful in describing how local morphological properties of a branch predict the downstream dendritic structure. Where relationships are found, new hypotheses can be generated with regard to both growth mechanisms as well as function. Yet other models disregard growth and focus instead of emergent functional properties by using techniques such as graph theory or evolutionary algorithms. **Models of Growth and Morphology** will focus on some of the techniques used in these various types of models and on the discoveries to which they have led.

Quantitative Measures and Functional Implications

Quantitative morphological measures, also called morphometrics, are the cornerstone of any morphological analysis. Such measures are numerous and many partially overlap each other in their meaning. They can be used to explore the differences between neuron classes, seek correlations with measures of function, and test hypotheses concerning development. They are also useful for growth models, specifically in comparing real neurons to growth models meant to simulate them, or as a basis for statistically based growth models. Measures of whole neurons or the separate domains of

axons, dendrites, and sometimes apical dendrites, can be broken down into various categories. Here we discuss them in terms of geometric, size-related, and topological features. Local geometric features, such as taper rate (

Figure 2.3a), tortuosity (

Figure 2.3b), and angle of bifurcation (

Figure 2.3c), are particularly useful for understanding influences on growth. Size measures include total length, surface area, and internal volume, as well as height, width, and depth (determined by a minimal bounding box [

Figure 2.3d] or by principal component analysis). There are also local size measures such as segment length and diameter. Topology deals with bifurcations, their numbers and the distribution of their resulting segments (i.e. connectivity). Topological features, as well as features that are combinations of topology and size, also fall into the category of morphological complexity. Descriptions of these measures along with some common uses and interpretations follow.

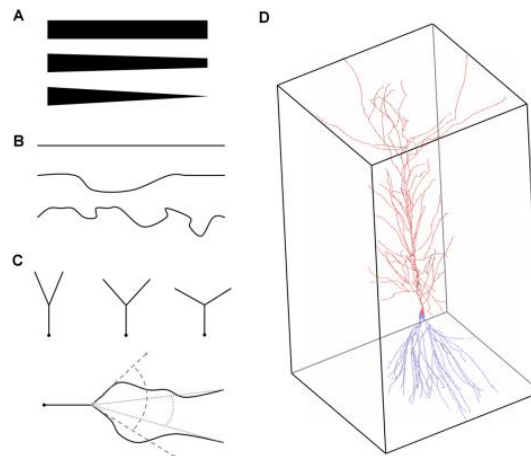


Figure 2.3. Size and geometry features. **a.** Taper rates from low to high (*top to bottom*). **b.** Increasing tortuosity (*top to bottom*). Tortuosity is the ratio between the path length of a segment and the Euclidean distance between its two ends. **c.** *Top:* Increasing angles of bifurcation. *Bottom:* Such angles can be measured locally as the angle of departure (*dashed lines*) or remotely as the angle between nodes (*dotted lines*). **(d)** The *bounding box* of a pyramidal neuron's dendrites (cell from Ishizuka *et al.*, 1995, via NeuroMorpho.Org).

Size and Geometry

A number of size and geometric features affect electrophysiological properties at local and global scales. Electrophysiological properties fall into the two categories of passive and active. Passive properties refer to basic resistance and capacitance effects on signal propagation, generally described by Cable Theory (Rall, 1969). Active properties refers to gated channels that change conductance depending on voltage resulting in nonlinear effects, usually spikes (i.e. action potentials), and are generally described by Hodgkin-Huxley equations (Hodgkin and Huxley, 1952). In experimentation, current can be injected at a constant level in a condition called current clamp (CC), allowing for the measurement of voltage. Alternatively, with a voltage clamp (VC) current is injected at a variable level in order to hold the voltage at a constant level. In all cases, a current input

(I) (via synaptic channels or an experimental patch pipette) will produce a change in voltage (V) across the membrane as described by the equation $V=IR$, where R is the input resistance. Capacitance describes the ability of the membrane to carry a charge such that a higher capacitance will result in more time for the membrane potential (i.e. voltage) to reach equilibrium in response to a change. Resistance and capacitance are both dependant on multiple morphological features, resulting in complex electrophysiological behavior within and between neurons.

Surface area is important to propagation and neuron function in that it is a primary determinant of a neuron's input resistance, the electrical resistance between the intracellular and extracellular space. A greater surface area allows for a greater number of channels through which ions/current can leak across, which lowers input resistance. In addition to decreased input resistance, increased surface area also results in increased capacitance. All else being equal, a cell with greater surface area will take longer to fire due to decreased input resistance and increased capacitance, but the capacitance will also store charge during a spike enabling a shorter delay before the next spike and therefore a higher maximal firing rate (**Figure 2.4a**). This relationship manifested in simulation as a lower burst rate (i.e. greater time between bursts) but higher within burst spike rate for larger compared to smaller CA3 pyramidal cell dendrites (Krichmar *et al.*, 2002). Greater surface area also allows for a larger number of synapses and therefore more possible connections to other presynaptic neurons, which has implications for structural plasticity and information storage (Chklovskii *et al.*, 2004; Stepanyants and Chklovskii, 2005).

Surface area is a product of a neuron's path length and diameter, each of which has additional functional implications.

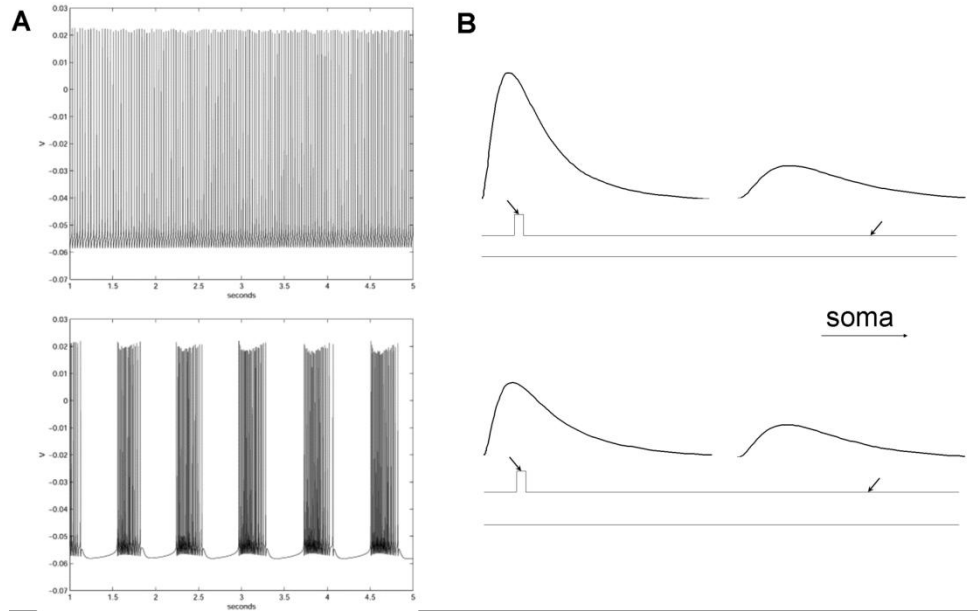


Figure 2.4. Morphological effects on signal propagation and firing rates. (a) A cell with less surface area (*top*) has a lower firing rate than seen within bursts of the larger cell (*bottom*). The faster within burst firing rate and gaps between bursts can both be explained by capacitive effects in concert with channel dynamics. (Adapted from [Krichmar *et al.*, 2002] with permission from Elsevier). (b) Voltage signals at location of synaptic input (*arrowheads* at left) and at a point closer to the soma (*arrows* at right) in two dendrites with differing diameters. The smaller diameter dendrite (*top*) has a greater input resistance and thus a larger peak voltage and faster rise time at the input relative to that seen in the larger diameter dendrite (*bottom*). The smaller dendrite also displays greater attenuation along the length due to the higher axial resistance. Flattening and delay of signal rise time due to capacitance can be seen between the signals in both dendrites.

Dendritic length plays several roles in neuronal function. Greater length tends to correspond to greater invaded space. This makes the dendrite available to more incoming connections. The tortuosity of a dendritic branch, or the ratio of the path distance between two consecutive branch points (or a branch point and a termination) and the Euclidean distance, provides information about the dendrite's growth mechanisms and its method of

reaching synaptic targets. This and bifurcation angles will be discussed further in **Models of Growth and Morphology**.

Electrotonically, greater lengths result in greater attenuation of a voltage signal given only passive propagation. This means that, all else being equal, synapses farther from the soma will have a smaller impact on somatic potential than those closer to the soma. In reality, there are a number of mechanisms by which a synapse can have a relatively increased effect, including optimized diameter properties, active dendritic propagation, and distance dependent synaptic strength. The attenuation based on increased length is due to input impedance which is the combined effects of the resistance and capacitance along the length of the dendrite. Capacitance is also responsible for signal flattening and delay, or the longer time to peak of a signal (**Figure 2.4b**). Rather than representing lengths in microns, particularly path lengths from synapse to soma, lengths are often represented in terms of the space constant λ , or the distance a signal travels before it is attenuated to $1/e$ of the initial voltage (Rall, 1969).

Along with membrane and cytoplasmic resistivity, diameter is the other major morphological factor in determining λ . While a larger diameter necessarily produces greater surface area and therefore locally decreases input resistance, it decreases axial resistance (i.e. resistance along the dendrite) by a greater amount. Axial resistance (per unit length of dendrite) is inversely proportional to the cross-sectional area of the dendrite, which is in turn proportional to the square of the diameter. Membrane resistance is inversely proportional to the circumference (and therefore to the diameter) of the dendrite. Capacitance is proportional to the circumference of the dendrite and is

responsible for the dendrite's input impedance and low-pass filter properties, though length effects are generally larger. In dendritic branches with greater diameter, the lower input resistance decreases the voltage achieved by a synaptic input, but the larger axial resistance better preserves the signal (**Figure 2.4b**).

The mismatch of impedance between the branches at a bifurcation is responsible for greater attenuation and increased probability of spike propagation failure (Vetter *et al.*, 2001). In order to avoid impedance mismatch at a bifurcation, the parent and daughter branch diameters must conform to the 3/2 power law, whereby the sum of each daughter branch diameter raised to the 3/2 power is equal to the parent branch diameter raised to the 3/2 power (Goldstein and Rall, 1974). The further away the relationship is from the rule, the greater the impedance mismatch and attenuation. This rule has been particularly useful for those who run computational simulations of electrophysiology. Where the rule holds (and assuming measurement of axo-somatic properties), bifurcations need not be individually represented and fewer compartments can be used.

The direct impact of dendritic diameter and length on electrotonic properties is seen in synaptic efficacy. One method for accurately measuring and visualizing these effects on synaptic efficacy given passive propagation is the morphoelectrotonic transform (MET) (Zador *et al.*, 1995). The MET visualizes signal attenuation and delay with an “attenogram” and a “delayogram”. Each neuron's visualizations are generated given an input site and an input waveform (e.g. a current pulse or a sinusoidal input). The calculations involved in the MET are similar to those for calculating electrotonic length. However, the MET produces an effective length constant (λ_{eff}) that does not assume

infinite cylinders as does the classical electrotonic length calculation. With the MET, one can view the electrotonic compactness not just from the soma, but from any individual potential synapse.

When considering active properties, length and diameter can have the exact opposite effect on synaptic efficacy as in the passive case (Komendantov and Ascoli, 2009). Simulations showed that higher input resistance at synapses farther from the soma lead to a greater probability of generating dendritic spikes therefore producing greater synaptic efficacy. Synapses closer to the soma with lower input resistance are relatively less likely to contribute to somatic spikes, i.e. they have lower synaptic efficacy. These two analyses on basic morphological properties and synaptic efficacy demonstrate the complex relationship between morphology, physiology, and function, yet there are still many more factors to consider.

Topology and Complexity

In addition to the various and important relationships between size and function, the distribution of those size attributes (i.e. where topologically and spatially length or volume are distributed) also has a vital impact on neuron function. The term “complexity” is often used to encompass measures of the distribution of a tree's size topologically or in space. Complexity measures do have functional implications, but many are more commonly used for classifying neuronal populations and validating models. A particularly common measure for complexity is the number of dendritic bifurcations, which can affect how signals are integrated in terms of impedance mismatch

(previously discussed) as well as how a neuron is integrated into its network in terms of space filling and synaptic connections (Wen and Chklovskii, 2008).

There are several terms used when discussing topology that are important to know. Many of the terms come from or have synonyms with graph theory, which is a system of describing the ways in which nodes can be connected and how one might find optimal paths given a set of nodes and edges (i.e. connections). A tree is a type of graph in which connections between nodes are unidirectional (i.e. directed edges) from parent to child and in which there are no cycles. In binary trees, nodes are either bifurcations, terminations, or the tree root. Degree refers to the number of terminal nodes in a tree, which is a useful indicator of tree size as it is directly proportional to the number of branches and the number of bifurcations ($branches = degree * 2 - 1$; $bifurcations = degree - 1$). A branch is the segment or edge between nodes and a branch's order (i.e. depth) refers to the number of branches between itself and the root node (**Figure 2.5b**). A branch's order is equivalent to the order of the bifurcating node that gives rise to it. Child and daughter branch are used interchangeably. These definitions will make it easier to understand topology and complexity.

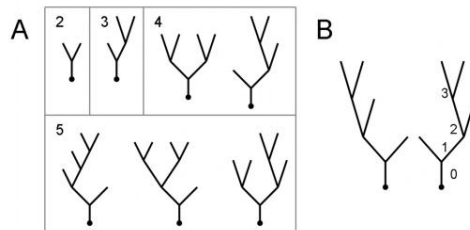


Figure 2.5. Unlabeled binary tree shapes. (a) All tree shapes of degree 2 through 5. (b) Two trees that appear different but are the same. They would be different if the nodes were labeled. The numbers refer to branch order. The root node and initial branch are of order 0.

Ignoring size and geometric features, and given a number of bifurcations, there are a specific number of unique binary trees called unlabeled (also referred to as unordered) binary tree shapes (Harding, 1971) as seen in **Figure 2.5**. While multifurcations do exist it is common to treat trees as binary. Any multifurcation can be broken into multiple bifurcations, though there are multiple ways to do so which can differentially impact the resulting topology. Methods are available to determine the most likely binary configuration for a multifurcation, assuming probabilities of the binary subtrees are known (Verwer and Van Pelt, 1990).

Vertex analysis was an early metric developed to analyze topology in neurons (Berry and Flinn, 1984). For binary trees the analysis focuses on the ratio of the number of “primary nodal vertices” (i.e. nodes with two terminating children) to the number of “secondary nodal vertices” (i.e. nodes with one bifurcating child and one terminating child). This method, along with asymmetry which is discussed next, was commonly used to determine the probability that a given tree or class of trees was the result of a particular growth process (discussed further in **Determining the Rules of Growth**).

Asymmetry is another property that is highly dependent on the mechanisms of growth. One particular measure of asymmetry, the tree asymmetry index, is measured at the tree level and is based on the partition asymmetry, where “partition” refers to the balance of size between daughter subtrees measured at a given bifurcation (Van Pelt *et al.*, 1992). The values of partition asymmetry are between 0 and 1 (inclusive) based on

the relative distribution (i.e. balance) of terminal nodes between a bifurcation's subtrees. A partition asymmetry value of 0 represents a symmetric distribution and a value of 1 represents a single termination in one subtree with the remaining terminations in the other subtree (**Figure 2.6a**). This version of asymmetry can also be referred to as topological asymmetry because it deals strictly with the distributions of terminal degree. Asymmetry can alternatively be calculated using distributions of length or other size attributes rather than degree. These other versions can be defined slightly differently because even in the case of a topologically asymmetric bifurcation the lengths of the branches are unconstrained. In the case of a node of degree 3 it would be possible for the side with the terminal node to have an equal or larger branch length than the combined length of the subtree of degree 2 (**Figure 2.6b**). The tree asymmetry index is defined as the average partition asymmetry across all bifurcations in a tree.

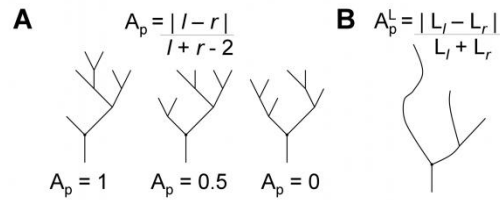


Figure 2.6. Asymmetry. (a) The partition asymmetry equation accompanied by trees (degree 6) with partition asymmetry (A_p) values of the first bifurcation node. The terms l and r refer to the degree of the *left* and *right* subtrees. (b) The equation for partition asymmetry by length accompanied by a tree that has a topological partition asymmetry of 1 but a length partition asymmetry of close to 0 due to the similar total length of *left* and *right* daughter subtrees.

Alternative versions of asymmetry can be produced by weighting the partition asymmetry values based on criteria depending on the purpose of the measure. Whether

one is interested in electrophysiological effects or differentiation of cell classes, if a particular property is suspected to be more greatly impacted by subtrees with greater size, one might choose to weight nodes by degree. Several alternative methods are raised in (Van Pelt *et al.*, 1992). The first is a scheme in which nodes of degree less than four are not considered. Such partitions are trivial with only one shape each. Another scheme weights nodes by the number of bifurcations in their subtrees (i.e. $\text{weight} = \text{degree} - 2$). The third alternative weights nodes by the number of possible distributions of bifurcations among the node's two subtrees (i.e. $\text{weight} = \text{degree} - 3$). Van Pelt and colleagues found that all schemes except for the second alternative displayed fairly large variances in tree shape samples and thus best served to differentiate different sets of trees from each other.

The impacts of asymmetry and topology were investigated in an experiment specifically focused on topology and its electrotonic effect. Van Ooyen *et al.* (2002) found that virtual dendrites with greater mean path length (of dendritic tips) had higher firing rates in response to somatic current injection. The experiment used all possible tree shapes with eight terminal nodes (i.e. of degree eight). Since branch lengths were all set equal to each other, the greater mean path lengths corresponded to trees with greater asymmetry. The tree asymmetry index did correlate with firing frequency, but not as well as mean path length did. However, branches closer to the soma have a greater influence on input resistance, so a weighted asymmetry would be a more representative metric and would be expected to better correlate with firing rate. The relationship between topology and firing rate ultimately is based on input resistance which rises (or input conductance

which decreases) with the increased mean path length of the more asymmetric trees. Another way to look at it is that a more symmetric tree has its branches connected in parallel, driving down input resistance, while a more asymmetric tree has its branches connected in series, driving up input resistance. A lower input resistance leads to more current leaking into the dendrites. However, increasing mean path length by increasing branch length has no impact on how the branches interact electrically. Surface area increases with the path length, which in turn decreases input resistance and therefore somatic firing rate.

Another variation on asymmetry, termed “excess asymmetry”, was used to determine whether the size of a dendritic branch or subtree is affected by the size of the rest of the dendrite. Excess asymmetry measures the extent to which a dendritic tree’s branches grow independently from each other (Samsonovich and Ascoli, 2006). One type of dependent growth would result in morphological homeostasis, in which a larger subtree would necessarily be offset by a smaller sibling. The excess partition asymmetry is defined as the difference between the partition asymmetry of a bifurcation and the average partition asymmetry of that bifurcation with its grandchild subtrees shuffled. The rationale is that, assuming morphological homeostasis, a larger subtree will be offset by a smaller sibling subtree and so shuffling the children of those subtrees will result in greater symmetry between the subtrees. Just as with the standard partition asymmetry, excess partition asymmetry can be calculated with regards to various size metrics. It was found that all sets of hippocampal CA1 and CA3 pyramidal cells tested had significantly positive excess asymmetry in terms of degree, length, and surface area. This result

complemented a result in the same study that showed morphological homeostasis between entire basal dendritic trees on the same cell, as well as between the basal dendritic arbor (i.e. all basal dendrites) and the apical dendritic arbor.

A topological property that is often visually evident is caulescence, or the prominence of a main path in a tree structure (Brown *et al.*, 2008). Like asymmetry, caulescence can be calculated using only topology or with any size attribute. Determining the caulescence of a tree is a two step process, the first being the determination of the primary branches (i.e. main path), the second being the determination of the value. In the first step, the primary path starts from a node (usually the soma or more generally a tree's root) and traverses away from the root choosing the larger daughter branch at each bifurcation (**Figure 2.7a**). The caulescence is the weighted average of the partition asymmetry along the primary path. As with the partition asymmetry calculation, any size metric can be used to determine which child is larger. It is customary to use the same size metric for both determining the path and calculating the partition asymmetry.

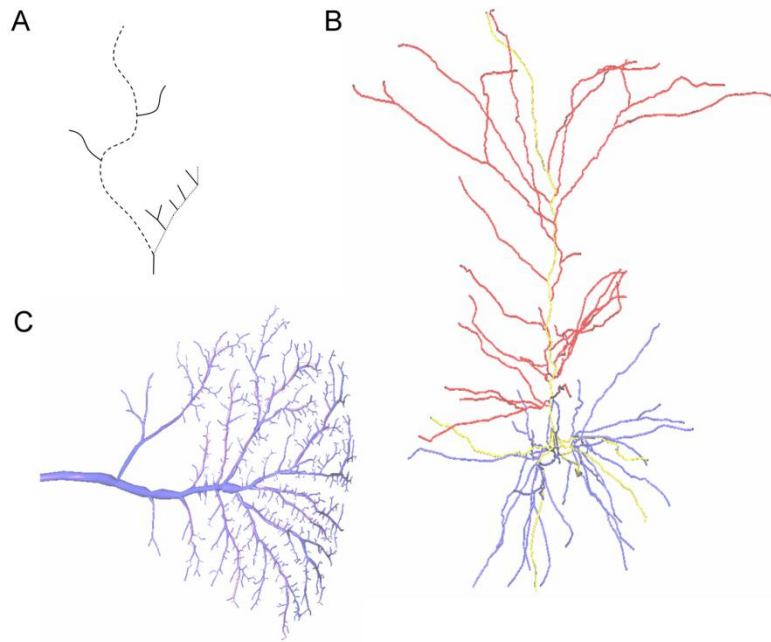


Figure 2.7. Caulescence. (a) Determination of the primary path in a tree by degree (*dotted*) and by total length (*dashed*). (b) Pyramidal cell dendrites (from Shepherd and Svoboda, 2005, via NeuroMorpho.Org) with the main paths by length (*yellow*) on apical (*red*) and basal (*blue*) dendrites. (c) Blowfly tangential cell (from Borst and Haag, 1996), via NeuroMorpho.Org) with visually approximated regions of high caulescence highlighted in *red*.

Axons are generally more caulescent than dendrites, but caulescence is also prominent in some classes of pyramidal cell apical dendrites (**Figure 2.7b**). Other cell classes may contain dendrites with multiple main paths or with main paths that appear in certain regions but do not start at the soma. Main paths are often found in dendrites in which synaptic targets are clustered at a distance from the soma, such as in the case of apical dendrites and the cluster of synapses in the apical tuft. A blowfly tangential cell dendrite appears fairly symmetric when considered as a whole, but between major bifurcations are segments that appear caulescent (**Figure 2.7c**).

Other complexity metrics that integrate topological and size attributes include fractal dimensions and Sholl analysis. Fractal dimension (Df) can be measured in several ways, each method producing an approximation of a form's true Df by performing some measurement at multiple scales (Jelinek and Fernandez, 1998). Fractals are common in nature, and while dendrites do not tend to strongly display fractal properties, fractal dimension can serve to differentiate different types of dendrites, including hippocampal neurons (Cannon *et al.*, 1999) and retinal ganglion cells (Caserta *et al.*, 1995). The latter experiment additionally showed correlations between fractal analysis and Sholl analysis. Sholl analysis (**Figure 2.8**) is a popular measure that charts the number of branches at set distances from the soma (Sholl, 1953). Originally these measures were done only in two dimensions as images of neurons were two-dimensional. With the advent to three-dimensional neuronal reconstructions, many metrics, including Sholl analysis, can now be carried out in three dimensions. Sholl analysis is often used to compare cell types or the effects of some experimental effect on a cell. It provides additional detail when size differences are found between two sets of dendrites. For instance, an experiment focusing on the effects of stress on neurons in the prefrontal cortex exhibited decreases in total length and number of pyramidal cell apical dendritic branches. Sholl analysis showed that within the 240 μ m that the dendrites reached, the reduced complexity occurred between 90 and 120 μ m of the soma, the region with the highest baseline complexity (most branch crossings) (Radley *et al.*, 2004).

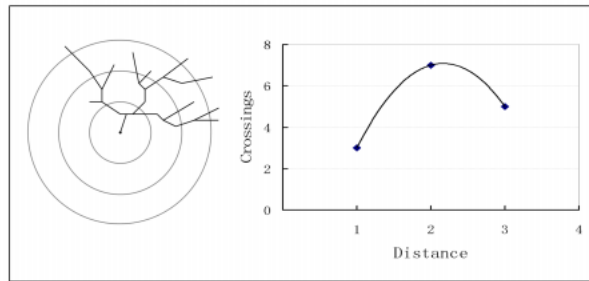


Figure 2.8. Sholl analysis design and plot. Concentric circles (*gray*) denote the distances from the center at which crossings of the sample dendritic tree (*black*) are counted. The graph on the right shows a sample Sholl plot based on the example tree. While Sholl analyses usually contain more bins to plot crossings, they do often have an upward then downward trend as shown.

Another metric that has come into more recent use is the tree-edit-distance.

Conceptually it is similar in function to the edit distance of strings; however trees contain a topology that strings lack. The distance is the number of insertions and deletions of nodes or branches required to turn one tree into a second tree. Nodes can additionally be labeled with values, perhaps features such as branch length or diameter. If such labels are used in an analysis, matched nodes must be modified (i.e. those not inserted or deleted), increasing the edit distance. The tree-edit-distance has been used to test whether neuromuscular projections are genetically determined (Lu *et al.*, 2009). Right and left side projections were matched by size within animals, and then distances were calculated. The distances were compared to distances calculated between neurons of different animals. If the neurons are genetically determined, one would expect the within animal differences to be smaller than the between animal differences, but this was not found. Using various node labeling schemes, the metric has also been used to classify several hippocampal neuronal types by clustering pair-wise distances (Heumann and Wittum, 2009). This metric is particularly interesting in that it only produces a difference between

two morphologies as opposed to providing a value for a single tree. Its benefit is the explicit comparison of topology merged with any particular combination of geometric or size attributes at the branch level.

Real Morphologies and their Properties

All of the discussed metrics can be used to compare dendritic trees in their entirety in order to classify them and understand the diversity within and between classes. A study by Cannon *et al.* (1999) provides a good example of a wide variety of size, complexity, and electrophysiological characteristics applied to the differentiation of several hippocampal cell classes. Multiple size characteristics sufficiently separated the interneurons, dentate gyrus granule cells, CA1, and CA3 pyramidal cells. Complexity metrics typically had more subtle differences with substantial overlaps in the distributions between classes.

One particular factor that should be kept in mind when using digital reconstructions for modeling or comparing neuronal types is the specific reconstruction methods used to develop the data. While standards and common practices are emerging, some aspects of the process, such as how the diameter of a dendritic segment should be determined, vary from lab to lab and even from person to person. These differences in procedure are in some cases due to resolution limits as well as the different scientific questions being probed. For instance, diameter could be traced at the inside of an irregular edge for the purposes of determining internal dendritic space for organelles to inhabit. Alternatively, tracing might follow the outside edge when considering potential synaptic connectivity based on distances from the dendrite. Assuming surface area to be a

key metric in a study, diameter could be traced in the middle to account for the surface area of untraced dendritic spines. As such, one cannot simply assume that all pieces of relevant information needed for a different question will be available. A study examined systematic differences between labs by comparing hippocampal CA3 and CA1 pyramidal cells with a battery of metrics (Scorcioni *et al.*, 2004). The two cell types were differentiated by several metrics, including remote bifurcation angle and maximum branch order, which did not vary consistently between labs. Total volume and diameter were the only two metrics that differentiated both cell types and labs. Local bifurcation angle and contraction (i.e. tortuosity) most differentiated labs, suggesting that they may be the most variable properties between labs. These results are useful for avoiding misinterpretation of certain metrics while designing or interpreting studies.

Morphometric analysis can also serve to describe and potentially differentiate interneurons, the roles of which are generally less clear than those of principal cells. Such an analysis, run on both dendrites and axons, successfully segregated a set of hippocampal CA3 interneurons into two novel interneuronal classes (Ascoli *et al.*, 2009). Initially distinguished by the number of dendritic bifurcations, further analysis found differences in a variety of local branch properties including branch length, taper, bifurcation angle, and asymmetry. The morphological findings combined with physiological data further suggested feed-forward and lateral inhibitory roles. Specifically, the interneurons target lateral pyramidal cells in CA3. Signals from upstream regions including the dentate gyrus and entorhinal cortex therefore produce feed-forward inhibition on the pyramidal cells. Recurrent collateral inputs from local

pyramidal cells onto the interneurons produce lateral inhibition on other nearby pyramidal cells targets, potentially enabling pattern separation.

Complete dendritic reconstructions exhibiting the full complexity of morphological influences are also used frequently in electrophysiological simulation. Simulation studies focusing on channel properties are quite common. With many different channel types exhibiting nonlinear responses to physiological state, there is a wide range of research on the topic that is beyond the scope of this chapter. However, some studies using reconstructions have focused specifically on the effect of morphology by keeping channel properties constant. By controlling for ionic channel type and distributions within dendrites, Mainen and Sejnowski (1996) found that various real morphologies in simulation were able to generate a spectrum of firing patterns which included weak adaptation, strong adaptation, and bursting behavior (**Figure 2.9**). These effects were further simulated in two compartment models by varying the coupling strength (i.e. the inverse of the resistance between the compartments) and area ratio between the dendritic and axo-somatic compartments. Some coupling was necessary to achieve bursting, but small changes to either variable had large impacts on cell behavior. Decreasing the area ratio or increasing the coupling strength reduced bursting. Changing channel properties can also alter firing properties, but these results showed how morphological properties alone can have a dramatic impact.

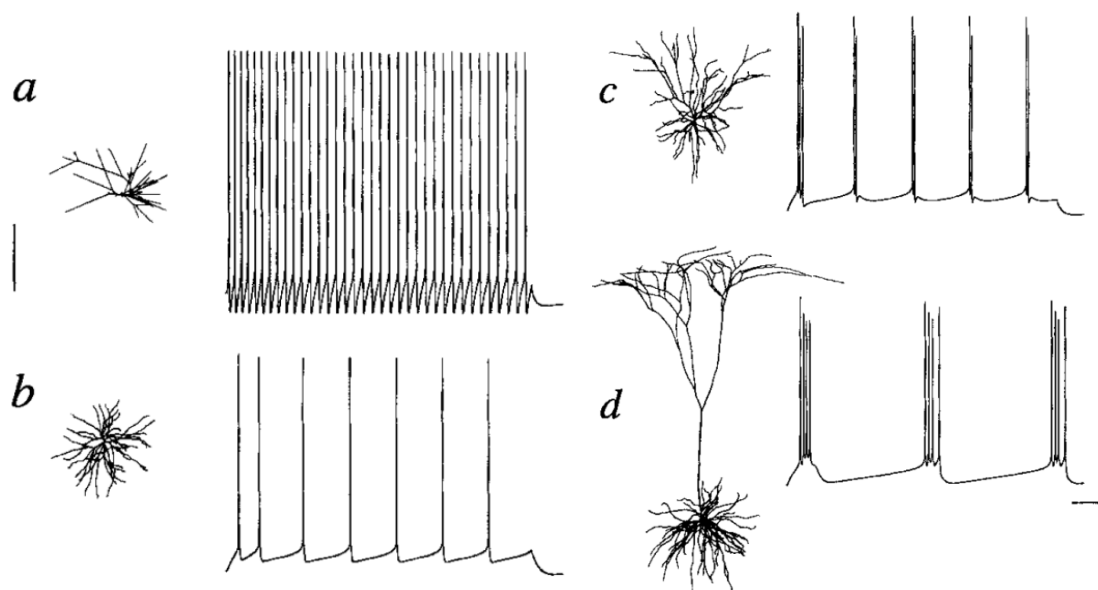


Figure 2.9. Morphology and electrophysiology. Simulated firing patterns produced by somatic current injection of several types of reconstructed dendrites (A shows only connectivity and branch lengths). Channel types and distributions are constant, only morphology is different between simulations. **(a)** Rat somatosensory cortex layer 3 aspiny stellate. **(b-d)** Cat visual cortex. **(b)** Layer 4 spiny stellate. **(c)** Layer 3 pyramidal. **(d)** Layer 5 pyramidal. Scale bars: 250 μm (anatomy), 100 ms, 25 mV. (Adapted by permission from Macmillan Publishers Ltd: Nature [Mainen and Sejnowski, 1996]).

These studies exemplify the importance of morphology in terms of neuronal function as well as morphometrics for producing hypotheses and providing evidence for such function. Size, geometry, and complexity each are relevant pieces of information. In addition to their roles in correlation with functional attributes and comparison between neuronal classes, they can be used as parameters in morphological models (discussed in **Models of Growth and Morphology**). The above section is far from an exhaustive list of metrics, but it contains the most fundamental and most widely used metrics along with some relatively new and intriguing entries in the field.

Molecular Underpinnings of Dendritic Development

With dendrites, as with all biological systems, genetic makeup and expression controls protein and molecular interactions. In conjunction with the external environment, these interactions generate a structure whose functional qualities and fitness shape the genetic makeup of successive generations through evolutionary pressures. Here we will discuss the genes and proteins that underlie dendritic development and therefore dendritic morphology. There are of course far too many interactions to cover in this chapter, so we will focus on a few examples of each type of interaction. More complete reviews can be found of transcriptional factors, molecular mechanisms, and signaling mechanisms (Miller and Kaplan, 2003; Gao, 2007; Parrish *et al.*, 2007). As many molecules and their interactions are modeled computationally in order to better understand dendritic growth, understanding a sample of the underlying interactions will in turn make it easier to understand the models and their results.

Cytoskeleton

Microtubules form the backbone of both dendrites and axons (Georges *et al.*, 2008; Conde and Caceres, 2009). In addition to providing structure, microtubules also act as tracks for the transport of resources, organelles, and signaling molecules by motor proteins in the dynein and kinesin families (Hirokawa, 1998). Initially, dendrites and axons both have microtubules with only plus ends distal. During early development, dendrites differentiate, carrying microtubules that face in both directions (Lewis *et al.*, 2009). The microtubules grow at their tail ends in the dendrites, with new tubulin proteins (the building blocks of microtubules) assembling distally into dendritic growth cones to stabilize them (**Figure 2.10**). The polarity of microtubules is due to the structure of their

alpha and beta tubulin subunits. As molecular motors selectively traverse microtubules in a particular direction, the mixed non-uniform polarity of dendritic microtubules allows for transport of a different set of proteins than seen in axons (Hirokawa and Takemura, 2005).

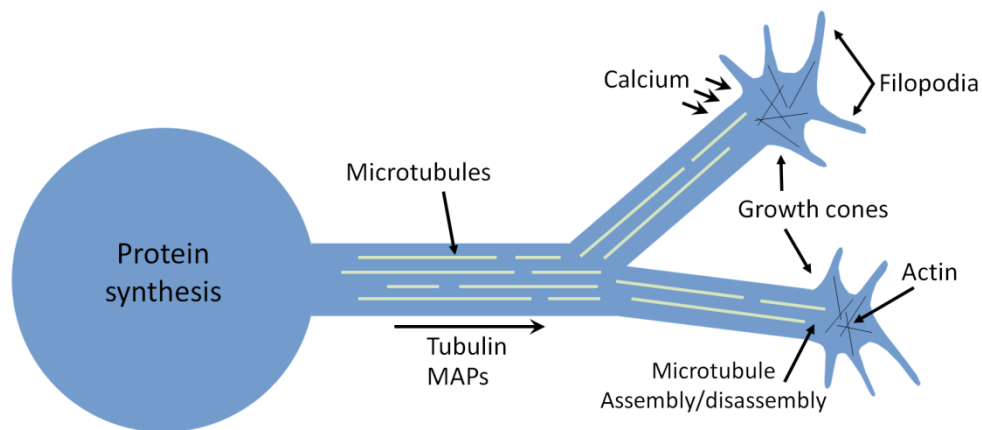


Figure 2.10. Developmental molecules and mechanisms of dendritic growth and branching. Tubulin and microtubule associated proteins (MAPs) diffuse and are transported from the soma to the growth cones. They interact to assemble microtubules and stabilize the dendrites as they elongate. The filopodia sample the extracellular environment, elongating the dendrite and causing bifurcations (i.e. branching) in some cases. Calcium influx influences microtubule stability and filopodia behavior.

Both microtubules and actin filaments provide stability to dendrites, but actin filaments also play important roles in dendrite growth and in the structure of dendritic spines. Like axons, dendrites have growth cones and filopodia composed of actin filaments that respond to the extracellular environment. Depending on the stability and spacing of the microtubules, strong growth signals in different directions can drive the filopodia to strain the microtubule bundles and split them into two groups, creating a

bifurcation. New branches can also form from existing ones in destabilized regions in a process called interstitial branching.

The stability of microtubules is determined in part by a family of proteins called microtubule associated proteins (MAPs). Some MAPs also associate with actin filaments, creating cross-links that further stabilize the entire cytoskeleton (Pedrotti *et al.*, 1994). MAP2 is specifically associated with the somatodendritic region in neurons. The protein increases assembly of microtubules as well as their stabilization via bundling, and it results in increased spacing between microtubules in dendrites relative to axons (Chen *et al.*, 1992). Phosphorylation of MAP2 is associated with reduced microtubule assembly (Yamamoto *et al.*, 1988) and with greater arborization of dendrites due to decreased microtubule stability (Díez-Guerra and Avila, 1993). MAP2 has many phosphorylation sites that have varying impact on microtubule interaction (Brugg and Matus, 1991), and there are many kinases and phosphatases capable of modulating its phosphorylation state. So controls over MAP2 expression as well as phosphorylation state provide mechanisms for control over dendritic growth and arborization. Calcium-dependent kinases and phosphatases specifically provide mechanisms for activity-dependent modulation of arborization, growth, and retraction (Quinlan and Halpain, 1996). While there are many other molecules and signaling pathways that influence actin and microtubule stability (Nikolić, 2008), the MAP family and their interactions provide a sufficient example of how dendrites can be effected to grow, bifurcate, and retract.

Regulatory Influences and Signaling Molecules

There are a great many intracellular features and signaling pathways that act both directly and indirectly on MAPs and the dendritic cytoskeleton. For example, in hippocampal neurons the Golgi apparatus (a specialized organelle in charge of post-translational modification and packaging of proteins) has been shown to orient towards the apical dendrite *in vivo* (Horton *et al.*, 2005). In a culture in which no neurite differentiated into an apical dendrite, the Golgi apparatus was oriented toward the largest dendrite. Moreover, Golgi elements were often found in proximal segments of larger dendrites. Horton and colleagues determined that the Golgi apparatus was responsible for the polarization by disrupting the Golgi apparatus organization without functional disruption. This resulted in a decrease in dendritic polarity without significant changes in total dendritic outgrowth. Another influence is the availability, or unavailability, of bifurcation inducing resources, as a function of dendritic volume and distance of travel via diffusion or active transport. That influence is a candidate for explaining how in some dendrites distal dendritic branches and particularly terminating branches tend to be longer than their more proximal counterparts (Graham and van Ooyen, 2004).

Dendritic growth is also guided and modulated by a number of different extracellular molecules and their interactions with receptors. Brain derived neurotrophic factor (BDNF), neurotrophin-3 (NT3), neurotrophin-4 (NT4), and nerve growth factor (NGF) are four major proteins of the neurotrophin family that impact dendritic growth by binding to specific members of the tropomyosin-related kinase (Trk) family of receptors and the p75 receptor. These growth factors have different influences depending on cell type, brain region, cortical layer, and even activity patterns (McAllister *et al.*, 1995), and

can sometimes have effects that counteract each other (McAllister *et al.*, 1997). These behaviors are made possible by the highly diverse signaling pathways triggered by the various Trk receptors, which include downstream effects on actin and microtubule dynamics (including MAP2) as well as transcriptional modulation of genes involved in dendritic growth. The term neurotrophin is sometimes used to refer to any protein that regulates neuronal growth. For instance, semaphorin3A, traditionally considered an axonal guidance molecule, also acts as an attractant for cortical apical dendrites (Polleux *et al.*, 2000). More neurotrophins and neurotrophic pathways are described in greater detail elsewhere (Donohue and Ascoli, 2005a).

Other transcriptional factors and molecules are important in stemming dendritic growth. The gene *sequoia* is at least in part responsible for tiling insect dendritic arborization neurons of the same class, preventing overlap of sensory receptive fields (Grueber *et al.*, 2002). Dendritic arborization neurons are also guided in part by the gene *Dscam*. Via alternative splicing in different neurons (Hattori *et al.*, 2008), the gene produces thousands of different cell surface proteins which enables dendritic self-avoidance (Hughes *et al.*, 2007). These and other related genes are vital for ensuring proper network connectivity.

Activity, or lack thereof, can also trigger growth and retraction of both spines and branches through numerous pathways, greatly impacting neuronal networks (Wong and Ghosh, 2002). In *Xenopus laevis* tadpoles, *in vivo* time-lapse imaging showed that optic tectal cells exhibited increased dendritic arborization (i.e. greater elongation and number of branches) when exposed to a lit environment relative to a dark environment (**Figure**

2.11) (Sin *et al.*, 2002). The effect was dependent upon a decrease in activation of the signaling protein RhoA and the activation of Rac and Cdc42. These Rho GTPases are involved in a number of activity- and neurotrophin-mediated pathways that influence dendritic and axonal growth (Luo, 2000; Miller and Kaplan, 2003). Activity during development of hippocampal pyramidal cells also increases dendritic growth, specifically close to the stimulated area (Maletic-Savatic *et al.*, 1999). In some systems, competition between inputs can drive extension and/or retraction of dendrites and axons. For example, activity during development drives multiple olfactory bulb mitral cell dendrites to retract leaving just one to interact with the olfactory glomeruli (Malun and Brunjes, 1996). While activity-dependent growth is often mediated by neurotransmitter activity, electric and magnetic fields can also influence the magnitude and direction of growth (Rajnicek *et al.*, 1992; Macias *et al.*, 2000).

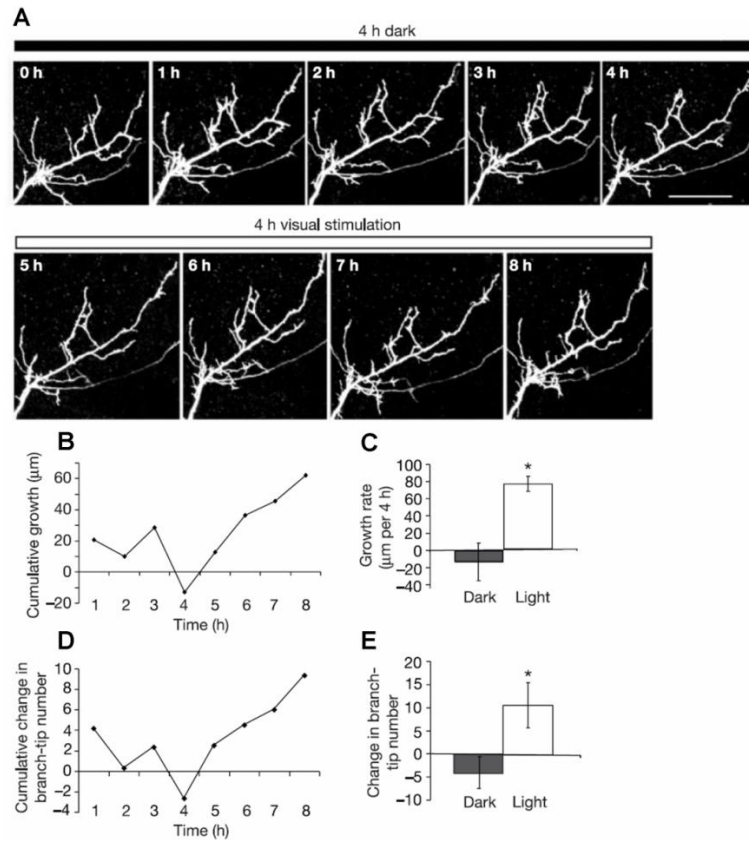


Figure 2.11. Example of stimulus induced growth. (a) Images of an optic tectal neuron taken with two-photon imaging, one per hour over 8 hours. The first image (0 h) is taken before light deprivation. Environment was dark during the first 4 hours, and a light stimulus was present for the latter 4 hours. Scale bar, 50 μm . (b) Average cumulative branch growth in terms of length. (c) Average growth rates during *dark* versus *light* periods. (d) Average cumulative change in terms of degree. (e) Average change in degree during *dark* versus *light* periods. (Adapted with permission from Macmillan Publishers Ltd: [Sin *et al.* 2002]).

Biologically Based Models

Clearly there are an enormous number of factors involved in dendritic growth and development. While it is the particular set of interactions that determine cell types, many factors can often be abstracted or simplified for the purposes of modeling. Aside from the difficulty in making sense of a complicated model where not all factors are fully understood, modeling all interacting factors would be impractical given current

computational power. Simplified models can provide insight into the dominant phenomena underlying morphology and help generate hypotheses for the function of proteins, pathways, and more complex mechanisms. Additionally, simplification allows for the isolation of the phenomenon of interest and can in principle ensure that an effect is a direct result of a particular manipulation.

Various biologically based models capture events such as neurite initiation, path finding, and elongation and bifurcation (Kiddie *et al.*, 2005; Graham and van Ooyen, 2006). One such example is a model of dendritic growth based on MAP2 concentration and phosphorylation and calcium concentrations (Hely *et al.*, 2001). The model used a constant rate of unbound MAP2 production in the soma, diffusion of MAP2 and calcium, and assumed tubulin availability. The MAP2 could become bound (i.e. associated with microtubules) based on rate equations, and could be phosphorylated and therefore dissociated from microtubules (**Figure 2.12a**). The phosphorylation and dephosphorylation rates were partially determined by calcium concentration. In addition to diffusing, the calcium had an influx component determined by each compartment's surface-to-volume ratio, resulting in greater influx in compartments of smaller diameter. At each time step, a terminal compartment would elongate at a rate proportional to the concentrations of bound and phosphorylated MAP2. Additionally, the terminal compartment had a probability of bifurcation based on the bound and unbound MAP2 proportions. In order to generate different morphologies, parameters affecting the reaction rates and sigmoid slopes of the phosphorylation and dephosphorylation functions were varied.

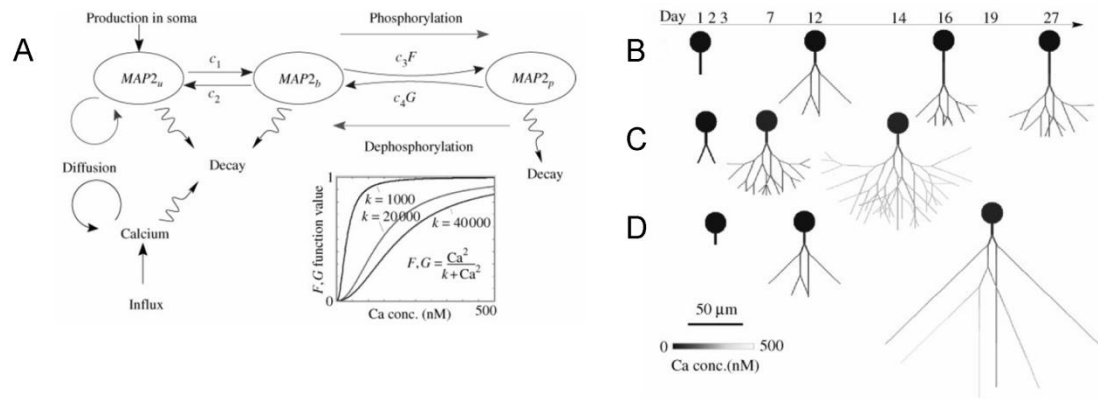


Figure 2.12. Modeling molecular-driven growth. (a) Molecular interactions modeled by Hely *et al.* (2001). (b-d) Dendrites generated with the MAP2 phosphorylation model with different parameters at various times during development. (b) *Low* initial branching probability, increasing over time/length. (c) Nearly constant branching probability and elongation rate. (d) *High* initial branching probability, decreasing with time/length. (Adapted from [Hely *et al.*, 2001], with permission from Elsevier).

Hely and colleagues produced three different types of morphologies with different parameters. The first type showed increased branching probability and decreased segment length towards the distal ends (**Figure 2.12b**). This was a result of a greater dephosphorylation to phosphorylation ratio due to the parameters strongly favoring dephosphorylation at low calcium concentrations. Calcium concentrations rose farther from the soma resulting in increased phosphorylation and branching. The second model had equivalent phosphorylation and dephosphorylation parameters, resulting in a more or less constant ratio of bound to phosphorylated MAP2 from soma to the distal ends (**Figure 2.12c**). This model also showed shorter terminal segments, which in both cases could be explained by a decrease in available MAP2 farther away from the soma and diffused throughout a greater total dendritic volume. The last parameter set was intended

to produce dendrites similar to pyramidal cell basal dendrites, with few bifurcations and much longer terminal branches than internal branches (**Figure 2.12d**). In this case the parameters were nearly the opposite of the first case, with phosphorylation being more dominant at low calcium concentrations. This led to more branching near the soma and less in the more distal region. The smaller volume as a result of fewer total branches allowed sufficient MAP2 for long distal branches. The resulting dendrites compared well to real basal dendrites in terms of distributions of degree, asymmetry, and several length properties.

As mentioned, most of the individual molecules and explicit mechanisms are not modeled when generating “virtual” (i.e. computationally derived) dendrites or understanding why dendrites take the shapes they do. In the example above, a small set of properties were modeled and showed how changes in calcium concentration in the context of possibly genetically encoded (de)phosphorylation set points could influence morphology. More generally, the emergent patterns of growth the many molecules and their interactions produce can be used in models to test the influence and relationships of various constraints, functions, and statistical distributions on morphologies of the many neuronal classes.

Models of Growth and Morphology

The relationships between various morphological and functional properties along with underlying mechanisms of growth and development are just a few pieces of the story of dendritic morphology. A wide variety of models are vital to elucidating how the growth mechanisms interact and how morphological and functional properties are

balanced to form different morphologies. Some models mathematically describe and simulate mechanisms of growth based on biophysical theories, either with molecular models as discussed in the previous section or by more abstract relationships between microtubules, dendritic diameter, elongation, and branching (Hillman, 1979; Tamori, 1993). Other models focus on the results of imposed constraints on growth, such as spatial boundaries (Luczak, 2006) or minimization of wiring and synaptic path length (Chklovskii, 2000; Cuntz *et al.*, 2007). Still others are statistical in nature, using empirical data to describe the relationships within and between fundamental and emergent features of various types of dendrites (Burke *et al.*, 1992; Ascoli and Krichmar, 2000; Samsonovich and Ascoli, 2003; Donohue and Ascoli, 2008). Not surprisingly, the boundaries between the different types of models are not always clear and many can be considered to fall into multiple categories. After all, it is ultimately the hypothesis to be tested that determines the design requirements for a model.

Determining the Rules of Growth

In an effort to better determine how dendrites grow in terms of where and when bifurcations occur, various models have been developed in which segments are stochastically added to a tree based on certain topological rules. It was proposed that such models could help determine how real dendrites and axons branch by comparing emergent properties of real trees with the emergent properties of trees grown from the models (Berry and Bradley, 1976). Since these models work by iteratively adding segments, the total size of the resulting tree could be controlled, thus enabling generation of virtual trees with the same size as real dendrites. Verwer and van Pelt explored the

statistical properties of trees generated by terminal growth (i.e. branches added only to terminating branches) or by segmental growth (i.e. branches added to terminating or bifurcating branches), seen in **Figure 2.13**. It was found that trees of much greater asymmetry and with a greater number of fully asymmetric nodes (i.e. nodes with one terminal child branch) were produced by segmental growth than by terminal growth (van Pelt and Verwer, 1983).

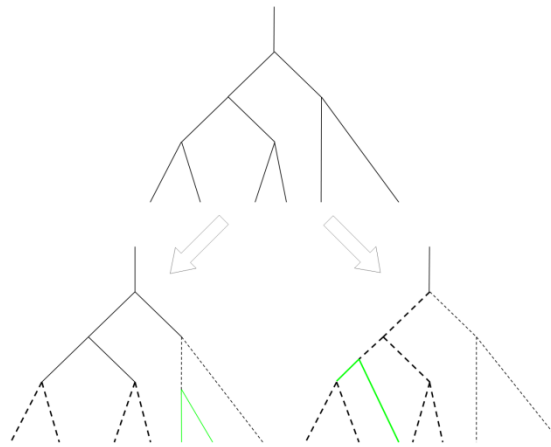


Figure 2.13. Terminal and segmental growth. Green lines represent the two new branches generated by the growth event (i.e. bifurcation) in the terminal (left) and segmental (right) growth cases. Dotted lines show all of the branches from which the new branch could emerge. Segmental growth gives a greater opportunity for the new segment to grow on the *larger left side* of the tree (*bold*) relative to the *smaller side*, with a 7:3 ratio, while the terminal growth ratio is 2:1. This results in greater asymmetry on average for segmental growth.

Subsequently, the Q-model was developed which provided a single parameter (Q) for controlling the relative probabilities of segmental and terminal growth. In order to integrate the observation that branching probabilities appear to often change with branch order, the S-model was developed. This model included order-dependent branching probabilities defined by the equation $2^{-S\gamma}$, where γ is branch order. Thus, S values less than zero produce increasing branching probability with increasing branch order, while

for values greater than zero branching probability decreases with increasing branch order (van Pelt and Verwer, 1983). The range of parameters allows for possible influences such as the greater space available for growth further from the soma, the greater availability of resources closer to the soma, and microtubule associated interactions such as MAP2 phosphorylation and others described in **Statistical Models**. The two parameter model combining the previous two models is referred to as the Q-S model.

The various models were used in conjunction with the tree asymmetry index to test growth hypotheses for several types of dendrites, including basal and apical dendritic trees and multipolar non-pyramidal dendrites (van Pelt *et al.*, 1986). When exploring terminal versus segmental growth alone, random terminal growth was insufficient to explain the asymmetry measures of half of the sets of dendrites, while purely segmental growth could explain none of the sets. Q model optimization resulted in a combination of segmental and terminal growth that fit four of the twelve sets of dendrites. However, all sets could be fit by optimized S and Q-S models, highlighting the importance of branch order dependence on branching probabilities. For instance, the distribution of basal dendrites from cortical pyramidal cells was best matched by Q and S values representing primarily terminal growth and a decreasing probability of bifurcation at higher branch order.

Another important factor in dendritic development, plasticity, and pathology is pruning (i.e. the loss/retraction of terminal branches). Applied to the previous models, various forms of pruning had different results (van Pelt, 1997). While uniform random pruning on trees grown under random terminal growth had no impact on tree asymmetry

index distribution, either uniform random pruning on trees grown with order-dependent branching or non-uniform (i.e. order-dependent) random pruning paired with the terminal growth model resulted in changes to asymmetry index distributions. A set of rat Purkinje cells taken at various ages during periods of growth and retraction was analyzed using these models. The cells showed no change in tree asymmetry index at the different time periods and sizes. The Q-S model parameters that best matched the cells suggested primarily terminal growth with no order dependence. Combined with the results of the pruning model, this suggests that the retraction process was a random uniform pruning process.

Other models have been developed to encompass a wide range of properties in addition to topology and branching probabilities. Hillman (1979) put forth an early version of the concept of generating dendrites entirely from a set of fundamental parameters specific to a given class of dendrite. In this case those parameters were determined based on observations of microtubules, diameter, and other morphological properties. The theory proposed that a dendrite with an initial diameter would elongate to some length with a particular taper rate, and branch if the diameter was above a particular threshold. The daughter branch diameters would then be determined by Rall's $3/2$ power law and a daughter branch ratio parameter. From a biophysics perspective, the relationships between bifurcation angle, diameter, and branch length can be calculated based on the principle of least effective volume, providing fairly accurate predictions of measured relationships for cortical pyramidal cell basal dendrites (Tamori, 1993). Both

formulations provided a framework for further development and application (discussed in **Statistical Models**).

Constraints and Functions

Like all biological systems, dendrites must balance a number of competing priorities under environmental constraints in order to contribute successfully to the neuron and its network. Their success depends on the interaction of their various features. Some priorities are known in principle, such as minimizing metabolic cost and integrating signals from synapses to the soma, but how those priorities are met and balanced is still an open question. A more diverse set of models have been used to answer questions relating to the influence of developmental constraints and emergent function. These questions can be answered by generating virtual dendrites in various ways, constraining either the growth or the final product based on some hypothesis, and comparing the result to real dendrites.

Environmental boundaries act as one of the major constraints on neurons. One particular growth model focused on the effect of generic growth factors within a defined space (Luczak, 2006). Luczak hypothesized that the shape of the environment along with the density and origin of neurotrophic factors could determine the morphology of various types of neuron. At initialization, the model contained an initial seed segment (soma) and a set amount of particles (neurotrophic factors) either uniformly distributed or clustered in some location (i.e. layer). At each iteration the particles diffused randomly, eventually contacting the soma or dendrite and becoming a new segment (**Figure 2.14a**).

Physiologically this simulates a neurotrophic factor coming in contact with a dendrite,

being absorbed, and triggering new growth at the point of contact. The model included a parameter controlling the span of iterations during which new terminal segments had a 40% chance of being pruned. This turned out to be a necessary component for producing realistic dendrites. Initial conditions were created to generate a wide variety of dendrites including granule cells, pyramidal cell apical and basal dendrites, interneurons and Purkinje cells (**Figure 2.14b**). The bounding box dimensions generally determined global dendritic shape, while particle concentrations determined branching density.

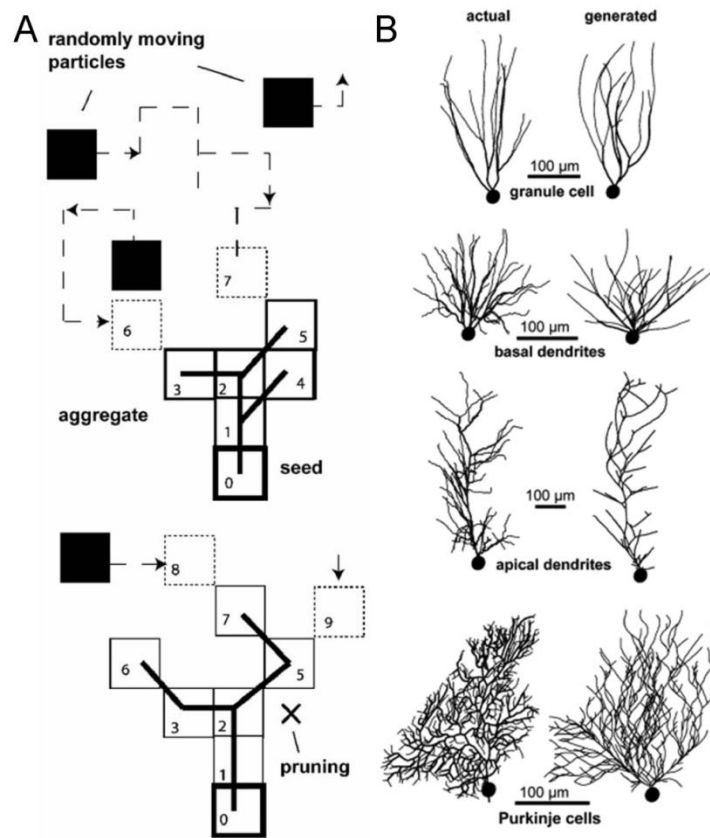


Figure 2.14. Modeling extrinsic factors. (a) Model schematic showing diffusion of particles, growth of the “aggregate”, and pruning. The numbers represent the order in which the particles were or will be converted into branches. (b) Real and generated dendrites of granule cells, and pyramidal cell basal and apical dendrites. (Adapted from [Luczak, 2006], with permission from Elsevier).

Multiple trees were grown in one space to observe the results of competition for particles and growth. Given sufficiently small spacing between seeds, the trees that happen to grow more quickly can then block other trees, resulting in larger trees getting larger and smaller trees staying small. Putting limits on growth prevented the problem. One downside to this model was the large size of the particles relative to the space. While computationally necessary, the large size of the particles can have an impact on where they are able to diffuse and prevents realistic generation of cells grown in close proximity. Nevertheless, this model does show that a set of fairly simple and primarily environmental constraints can produce a variety of morphologically distinct dendrites.

In an approach focused on intrinsic constraints, (Cuntz *et al.*, 2007) explored the extent to which dendritic morphology can be reproduced solely by minimizing wiring and optimizing synaptic efficacy by minimizing path distance. In order to generate a dendrite, their model took a series of nodes, derived from a blowfly lobula plate tangential cell (**Figure 2.15a**), and connected them based on graph theory concepts. An initial version of the model minimized only wiring with the minimum spanning tree algorithm, starting with the root node and connecting one node to the dendritic tree one at a time, adding the least wiring possible. The nodes were randomly sampled from the entire dendritic spanning field of a real tangential cell. The results were fairly unlike real dendrites, having some particularly long paths (**Figure 2.15b**). In order to minimize path distance in addition to wiring, the minimum spanning tree algorithm was made to consider the path distance of any potential new connection. A balancing factor was used to balance

minimization of wiring and path distance. The factor was optimized to produce qualitatively realistic tangential cells (**Figure 2.15c**).

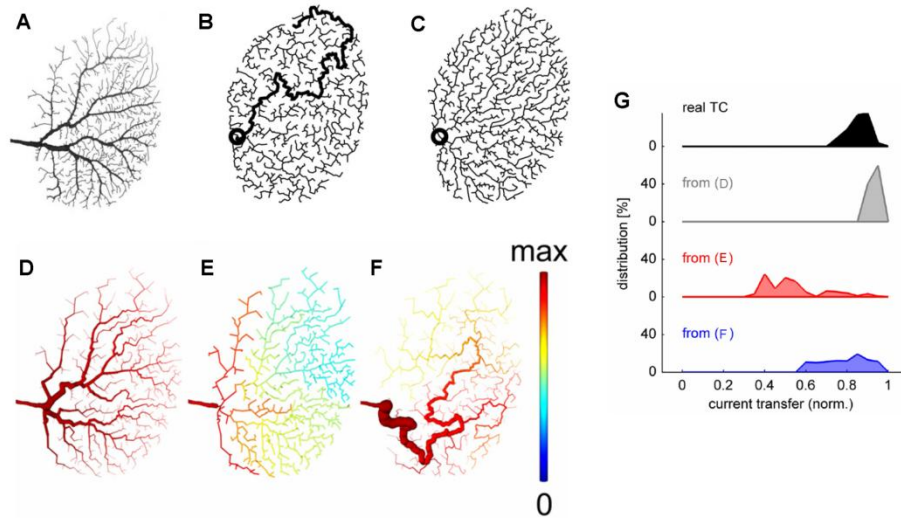


Figure 2.15. Functionally constrained modeling. (a) Original tangential cell dendrite. (b) Virtual dendrite generated by minimizing wiring and connecting randomly sampled points within the spanning field of the dendrite in (a). (c) Virtual dendrite generated by minimizing wiring and path distance using the same points used in (b). (d-f) Trees colored by current transfer to the root. (d) Virtual dendrite generated by minimizing wiring and path distance and connecting the bifurcation and termination points of the dendrite in (a). Diameter has been set to optimize synaptic efficacy. (e) The same tree as in (d), but all branches have equal diameter. (f) Virtual dendrite generated by minimizing wiring, using the same points as in (d) and (e). (g) Distributions of normalized current transfer from the real tangential cell and from trees in (d) through (f) show that minimizing wiring and path distance and optimizing diameter best reproduces the current transfer profile of the real cell (Merged selection of two figures from [Cuntz *et al.*, 2007], with permission from BioMed Central).

In addition to manipulating topology to optimize synaptic efficacy, Cuntz and colleagues altered dendritic diameter. The input resistance of a synapse increases with distance from the soma. Therefore a voltage signal at a synapse will attenuate more at the soma with increased distance given passive dendritic properties. Dendritic diameter profiles were therefore found which optimized the current transfer from any given location to the soma, minimizing attenuation. In order to focus on that particular attribute,

the wiring and path length minimization model was run on a set of nodes composed of all bifurcations and terminations of the real dendrite. **Figure 2.15d-g** show that the dendrite generated by the model minimizing wiring and path length and with the optimized diameters was most successful at optimizing synaptic efficacy and compared well to the real dendrite.

A similar conclusion was reached by Wen and Chklovskii (2008) for Purkinje cells by calculating mesh sizes (i.e. the ratio between a sample area and the total dendritic length within that area) for various types of branching schemes, minimizing wiring and path lengths, and avoiding overlap of potential synapses. Since the role of dendrites is to make synaptic connections, a dendrite's branches should be spaced sufficiently such that one branch does not invade the space in which another branch might form a synapse. The equations developed showed that Purkinje cells' mesh size fit within the theoretical calculations. Purkinje cells were used because they are approximately two-dimensional in shape and exhibited non-overlapping dendrites. These works suggest that efficiency through minimized wiring is a major pressure in determining dendritic morphology.

Another method that has been developed for relating morphology and function produces dendrites through an evolutionary process starting with a particular functional constraint (Stiefel and Sejnowski, 2007). With this method generated dendrites can be analyzed to determine the ranges and types of properties necessary to achieve the electrophysiological function, and then compared to real dendrites. The process developed involves parameter sets for creating dendrites, electrophysiological simulation,

fitness scoring of the simulation, and selection, mutation, and recombination to produce successive generations with increasing fitness.

The two functions proposed and tested in a proof of concept were linear summation of synaptic inputs and spike-order detection (i.e. A before B, but not B before A). The linear summation selection resulted in neurons with long, thin, and highly polarized dendrites with a large electrotonic separation. Thus, the input current of one synapse would have a minimal impact on the driving force of the other. As predicted, the dendrites looked very much like those of the crocodile and bird nucleus laminaris neurons as well as mammalian medial superior olive neurons, each of which receives auditory input from both ears separately onto polarized dendrites. The polarized and separated morphology of the dendrites enables reaction to simultaneous activation of two different sets of synapses but not to activation of just one set. The spike-order detection selection resulted in two sets of dendrites with different diameters. Thus one set of dendrites acted as strong low-pass filters while the other dendrites were much weaker low-pass filters. This relationship resulted in a greater somatic voltage when the strong low-pass filter dendritic input came first. The authors pointed out that this relationship exists in pyramidal cells between the apical and basal dendrites. Whether spike-order detection is actually used by pyramidal cells is an open question, though the presence of active properties makes answering that question more complicated. With parameters for active properties, synaptic distribution, and simulations of spiking properties, this type of evolutionary model could be useful for generating experimental hypotheses for a wide range of functions.

Electrophysiological effects and efficiency demands appear to be largely responsible for shaping dendritic structure given a spanning field for synaptic locations. Be that the case, the question still remains why synapses are located where they are. A possible component to the answer comes again from considering wiring efficiency, both in terms of total wiring in the brain (Chklovskii *et al.*, 2002) and in the trade-off between axonal and dendritic size (Chklovskii, 2000). Consider two sets of neurons with one set being an input (i.e. sending) layer and the other set being an output (i.e. receiving) layer. Now consider that there are twice as many input neurons as output neurons, 2 input (i.e. presynaptic) and 4 output (i.e. postsynaptic) as an example. Moreover, each input neuron sends out 2 axons; this is referred to as the “divergence” (D). Assuming each layer is more or less homogeneous, each output neuron will receive 1 connection. This is referred to as the “convergence” (C). In this case, where $D > C$, wiring is minimized by the axons being larger and growing to meet the dendrites of the input neurons (**Figure 2.16**). The mathematics used to prove this is based on the spatial density of pre and postsynaptic neurons in a one-dimensional scheme. In two dimensions and with larger numbers of neurons, neurite (i.e. axon or dendrite) “meshes” are used. Chklovskii showed that the optimal ratio of dendritic to axonal length is given by the square-root of the ratio of convergence to divergence. Factoring in differences in axonal and dendritic diameters, the optimal ratio is then adjusted by multiplying the previous ratio by the square root of the axonal to dendritic cross-sectional area ratio. Thus a larger axonal cross-sectional area would result in greater dendritic length and less axonal length. Anatomical data from a

variety of experiments on retinal, cerebellar, olfactory bulb, and neocortical neurons support the theory (Chklovskii, 2000).

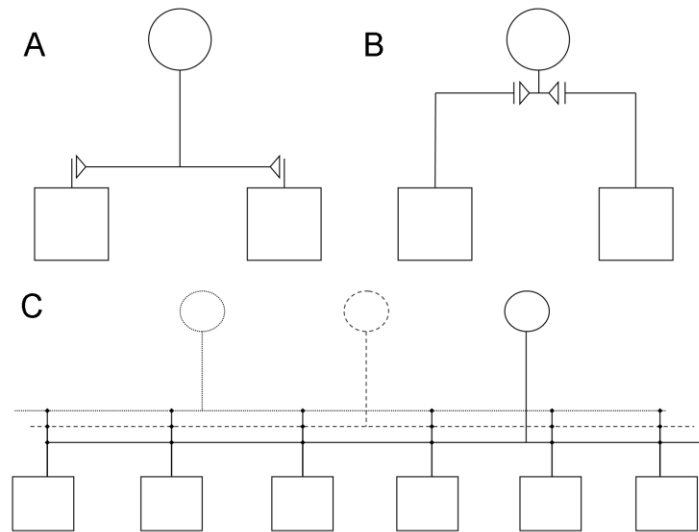


Figure 2.16. Divergence and convergence. With a convergence of 1 and a divergence of 2, the layout in (a) with a longer axon is more efficient than the layout in (b) with larger dendrites. (c) Example of complete and optimized connectivity with convergence of 3 and divergence of 6.

It was further shown that within gray matter optimization theory requires that axonal and dendritic wiring take up 60% (bounded from 50% to 75% depending on certain assumptions) of the volume (Chklovskii *et al.*, 2002). The calculations used independently minimized axonal conduction delay and dendritic attenuation, maximized the number of potential synapses, or minimized wiring length while in each case holding the other variables constant. For instance, conduction delays can be reduced by larger axonal diameter, but with sufficiently large diameters the total volume of the gray matter must increase and thus axons must travel farther. Data from neocortex, piriform cortex, and hippocampus show wiring percentages not significantly different from 60%. Thus the

evidence suggests that in terms of gross size attributes, axonal and dendritic morphology are results of evolutionary processes optimizing signal propagation (in terms of speed and fidelity), potential synaptic connectivity, and total wiring.

Statistical Models

While statistical models do not provide information on how or why dendrites grow into their various shapes, they are very useful for providing compact descriptions of various morphological classes. These models generally attempt to capture the statistical relationships between known morphological attributes of cell classes, including the variance in those attributes. This is often achieved by using a subset of “fundamental” properties obtained from neuronal reconstructions, such as diameter or branch order, to probabilistically determine other properties, such as length or likelihood of bifurcation or termination. While many models have “growth rules”, they do not necessarily simulate developmental growth but rather describe the relationship between parameters in (usually) adult cells in a branch-by-branch manner. One major goal of such models is to expand the store of available morphologies for more physiologically realistic network models. Given the intense amount labor that is currently required to reconstruct individual cells, an accurate statistical model can generate a larger set of realistic morphologies. Thus the need to store large amounts of data is reduced as the morphologies can be generated “on the fly” (Ascoli, 2002).

Burke et al. (1992) developed such a model in which branches iteratively bifurcated, terminated, or elongated with probabilities based on observed diameter relationships. Elongation length per iteration and taper rate were set as global parameters

while bifurcation and termination probabilities were determined by the current diameter of a segment. Daughter branch diameters satisfied the $3/2$ power ratio by an observed margin and were determined by a process that produced the same distribution as seen in the real dendrites. The statistical data determining the probabilities came from reconstructed cat spinal cord α -motoneuron dendrites. By adjusting the global taper rate, the dendritic trees generated from the model had realistic branch order distributions but unrealistic path distance distributions. The addition of path distance (from the soma to the growing segment) as a parameter (in conjunction with diameter) in determining bifurcation and termination probabilities was sufficient to produce realistic path distance distributions, though some discrepancies such as total surface area remained.

In order to enable the further study of various statistical and anatomical models, the program L-Neuron was developed (Ascoli and Krichmar, 2000). L-Neuron flexibly implements the concepts of Hillman, Tamori, and Burke, allowing an experimenter the option of using statistical distributions or biophysically derived calculations to generate morphologies from fundamental properties. The program additionally includes parameters that allow for path lengths larger than the straight distance between nodes (i.e. tortuosity) and global tropisms simulating external guidance or somatrepulsive effects. Moreover, the program can be fed statistical data of a given neuronal class or digital morphology files in order to generate that data. The models in L-Neuron were evaluated using motoneurons and Purkinje cells (Ascoli *et al.*, 2001). Models based on primarily local constraints, such as diameter, successfully produced virtual dendrites with several emergent parameters, including degree, total length, and asymmetry, which matched both

cell types. The global tropic influences improved results in terms of spatial distribution of the dendrites. In general however, variability in virtual dendrites was greater than real dendrites, and certain emergent parameters were better reproduced by certain models. The results suggested that further constraints were necessary and that some combination of the models might produce improvements.

Variations of the Hillman model were explored using L-Neuron for hippocampal CA1 pyramidal cells. It was initially found that constraining branching by diameter alone resulted in excessively large neurons or neurons that failed to stop growing, which was due to selection of taper rates (Donohue *et al.*, 2002). Rather than sampling a taper rate for the entire tree from the entire distribution of trees, a further variation on the model was developed in which taper rates for each branch were sampled based on the branch's diameter (Donohue and Ascoli, 2005b). The updated model successfully reproduced the size (i.e. degree) of basal dendrites. Apical dendrites were also less likely to grow out of control, but their sizes were still more varied relative to the distribution seen in real CA1 apical dendrites. In terms of branching patterns, the model actually captured apical dendrite asymmetry fairly closely, and better than it captured basal dendrite asymmetry.

Further work in this direction tested the abilities of branch order and path distance as determinants of local branch properties (i.e. taper rate, bifurcation probability, length, parent-daughter ratio, and daughter ratio) to generate realistic apical, basal, and non-pyramidal dendrites (Donohue and Ascoli, 2008). Each determinant produced the best results for some combination of dendrite type and emergent parameter (e.g. degree, asymmetry, surface area, and surface area asymmetry). Given the results, two models

were created that merged the influence of branch order, path distance, and diameter in the hopes that a combination of influences would reflect real statistical relationships and produce a more accurate model. The percent mix model merged the sampled values of the local properties with various weights, while “243 Mix” model assigned each local property with one of the determinants. Many weights were run for the percent mix model while for the “243 Mix” all determinant/local property assignments were run to determine the best combination of influence. After normalizing for the large numbers of weights and combinations sampled, the “243 Mix” model performed better than the percent mix or the individual determinants in terms of bifurcation asymmetry and surface area. The particular associations between local basic properties, their determinants, and the emergent properties may provide some idea of how growth mechanisms are determined in the different types of dendrites. For instance, diameter may be a better determinant of bifurcation asymmetry due to the fact that segmental branching tends to result in smaller side branches in terms of diameter and degree. Path distance may serve a role in determining the limits of both elongation and branching via resource availability and transport.

Several of the functional and constraints based models guide growth by primarily extrinsic factors; however there are several potential intrinsic factors that may also determine a dendrite's tropism (i.e. direction of growth). Hippocampal principal cell dendrites have a tendency to grow away from their somas and appear mostly contained within a conical region. Cytoskeletal rigidity may be expected to constrain changes in direction, however when a dendritic branch does turn it seems to usually return to its

original direction (**Figure 2.17a**). This provided the motivation to determine whether the dendrites are primarily directed by extrinsic factors promoting growth in a particular direction, by intrinsic factors promoting growth away from their soma, or by neither type of influence (Samsonovich and Ascoli, 2003). In order to make the determination a statistical model using Bayesian analysis was developed which represented the growth of non-branching dendritic segments by three vectors of influence with additional random deflection from the combined vector's direction (**Figure 2.17b**). The three vectors included a continuation of the direction of the previous segment representing cytoskeletal rigidity, a unidirectional influence representing some extrinsic factor(s), and a radial vector representing a somatorepulsive influence. Using reconstruction data, the magnitudes of the unidirectional and somatorepulsive influences were optimized to minimize the amount of random deflection. The Bayesian analysis showed that the radial tropism was significantly stronger than the fixed direction influence, and that its strength decreased with greater distance from the soma. Moreover, the radial growth was greater in the basal relative to apical dendrites. Similar results were seen in α -motoneuron dendrites but not γ -motoneuron dendrites when analyzing daughter branch orientations relative to parent branch orientation and bifurcation position (Marks and Burke, 2007). These results suggest a possible functional benefit of growth away from the soma for certain cell types.

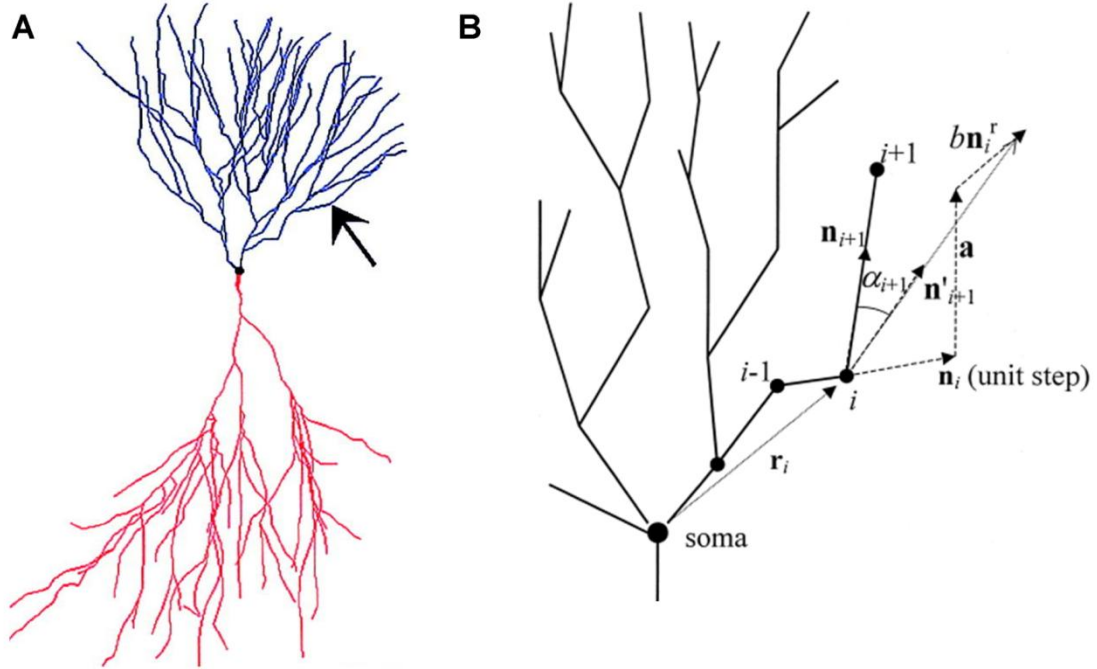


Figure 2.17. Modeling somatopulsion (a) Pyramidal cell featuring roughly conical boundaries. *Arrow* points out a clear instance of a turn towards a somatofugal direction. (b) Model schematic. Current node is i . The three vectors determining the direction of the next elongation are \mathbf{n}_i in the direction of the previous step, \mathbf{a} in the direction of the constant extrinsic influence, and $b\mathbf{n}_i^r$ in the direction away from the soma (\mathbf{r}_i). The angle α_{i+1} randomly changes the next step, which determined by the three vectors is \mathbf{n}'_{i+1} , to the actual next step \mathbf{n}_{i+1} . (Adapted from selected figures in [Samsonovich and Ascoli, 2003], with permission from John Wiley and Sons).

Each type of modeling has clear utility as well as limitations. Mechanistic models that include molecular concentrations and interactions are usually computationally expensive. Though good at showing how changing certain basic parameters can influence emergent properties, most details are left out. For instance, in order to be practical they restrict growth to one or two dimensions (Kiddie *et al.*, 2005). Luczak's model using neutrophin-like particles constrained by boundaries showed that those factors could largely explain how many types of dendrites might grow, but due to both computational constraints and a lack of additional interacting factors the results were only

approximations of real morphologies. Function-driven models rarely grow dendrites in a realistic manner as their goal is to relate completed morphologies with functional capabilities. Statistical models can serve to bridge the gap between mechanistic and function-driven models. They can account for various molecular interactions and activity effects as variance in the model, though this also means that any error in the experimental data or limitation of the reconstructions will also be captured in the model. Moreover they produce quantitative values of the relationships between morphological attributes which ultimately represent the range of morphologies that support the particular function of the dendrite being modeled.

Dendritic Morphology into the Future

There is still room for exploration of the relationship between morphology and electrophysiology, but many aspects have been investigated in great depth. Several of the models discussed focus on the relationship between network connectivity and morphology, of which much more is yet to be discovered. As time passes, molecular neuroscientists will continue to discover pathways and relationships between various growth factors, signaling molecules, and dendritic growth and plasticity. Some may be applicable to all dendrites, but a number will be specific to certain cell classes based on genetic expression. More importantly, many of those will be specifically involved in setting up connections between various anatomical regions and functional networks.

Researchers at the Allen Institute for Brain Science (2009), producers of the Allen Brain Atlas (a map of gene expression patterns in the mouse brain [<http://www.brain-map.org>]), found gene expression differences in sub-regions of the hippocampus

(Thompson *et al.*, 2008). Cell adhesion genes made up a large proportion of the genes that produced the distinct boundaries, suggesting that the genes are at least in part responsible for hippocampal circuitry. Previous tracing studies have provided evidence for some of these regional divisions, but these new results provide an opportunity to further explore the relationships between genetic expression, molecular growth mechanisms, morphology, and network connectivity in very specific subregions.

It is clear that the region of synaptic targets of a neuron play a role in determining the neuron's dendritic morphology. At least one cell class's morphology can be fairly accurately reproduced with knowledge of only the spanning field while minimizing both wiring and synaptic distance to the soma (Cuntz *et al.*, 2008). In fact the general cell class of lobula plate tangential cells is composed of four subclasses that the referenced method suggested were solely differentiated by their spanning fields. Whether this model extends to all classes of dendrite is unclear. Might the balance between minimizing wiring and synaptic difference be shifted in one direction or another in other dendrites? Even if that question is answered, the question of what determines the shape and location of the dendrites' spanning fields remains. As discussed earlier (**Constraints and Functions**), Chklovskii's models and equations focused on minimizing wiring in the brain suggest one likely component to the answer.

Dendritic morphology is shaped by many forces, both in terms of development as well as function. Varieties of intrinsic genetic, molecular, and electrophysiological factors interact with extrinsic spatial constraints, neurotrophic factors, and network activity to generate the neurons we observe. All of those interactions are influenced by

evolutionary pressures to achieve the electrophysiological functionality and network connectivity that enables a neuron to take part in the local and global networks that make up the brain. With advances in imaging such as the individualized cell labeling of Brainbow (Livet *et al.*, 2007) and potential automated reconstruction algorithms, we can expect an explosion of morphological and connectomic data. It will then take the confluence of database management, morphometric analysis, and the broad array of modeling methods to bring us closer to a complete understanding of the many relationships dendritic morphology has with brain function.

Acknowledgements

We thank Michele Ferrante and Kerry Brown for their valuable discussions and feedback.

CHAPTER THREE: TOWARDS HIGH-THROUGHPUT DATA PROCESSING BY PROMOTING DEVELOPMENT OF AUTOMATED MORPHOLOGICAL RECONSTRUCTION

3.1. DIADEMchallenge.Org: A compendium of resources fostering the continuous development of automated neuronal reconstruction

[With kind permission from Springer Science+Business Media: *Neuroinformatics*, volume 9, 2011, Todd A. Gillette, Kerry M. Brown, Karel Svoboda, Yuan Liu, Giorgio A. Ascoli, © Springer Science+Business Media, LLC 2011]

Concomitant with the publication of this Special Issue of Neuroinformatics, a substantially updated version of the DIADEM web site has been released at <http://diademchallenge.org>. This web site was originally designed to host the challenge for automating the digital reconstruction of axonal and dendritic morphology (hence the DIADEM acronym). This post-competition version features additional content for continued use as the access point for DIADEM-related material. From the very beginning, one of the spirits of DIADEM has been to share data and resources with the neuroscience research community at large. The resources available from or linked to the DIADEM website constitute a substantial scientific legacy of the 2009/2010 competition. The new content includes finalist algorithms, image stack data, gold standard reconstructions, an updated DIADEM metric, and a retrospective on the competition in text and images. The website continuing intent is to facilitate development of automated reconstruction algorithms.

The DIADEM Data Sets include image stacks, manually reconstructed digital tracings (the “gold standards”), and metadata. The six extensively curated, diverse data sets can be used to train, test and aid in the development or tuning of existing automated reconstruction algorithms. The previously posted image stacks (training and qualifier sets for the DIADEM competition) are still available, and are augmented with the addition of the Final Round image stacks (Brown *et al.*, 2011). The Final Round sets include a previously unreleased type of data, a visual cortical pyramidal cell, used as a surprise set in the competition.

The DIADEM metric (including source code) also has a new release with a wider set of user options and extended documentation (Gillette *et al.*, 2011b). These options provide flexibility for use on data beyond the DIADEM data sets and for alternative scoring schemes. Integration of the metric with automated reconstruction development has been made easier with more output options and methods to make results accessible to an interfacing program. These and other changes make the DIADEM metric viable in both algorithm development and evaluation. Since it was designed as a machine-derived surrogate to humans for quantifying the differences between two reconstructions of the same neuronal structure, the DIADEM metric can be used as a reliable tool to evaluate reconstruction quality of either newly developed or expansions of existing automated reconstruction algorithms. Most importantly, the DIADEM finalists made their algorithms freely available on the website. These include code, executables, and documentation, in addition to links to the development web sites and authors’ contact information. Several additional resources are also accessible from the DIADEM web site

links, including databases of digitally reconstructed neurons and dynamic inventories of web-based neuroscience knowledge, data, and materials, as well as 3D morphology conversion utilities, freeware software for neuronal reconstruction and editing, open source tools for online neuromorphometric analysis, and visualization programs.

The DIADEM web site and its resources, in conjunction with the DIADEM listserv (http://listserv.janelia.org/mailman/listinfo/diadem_challenge), will foster the increased collaboration required to speed up development of the next generation of automated neuronal reconstruction algorithms. Ultimately the aim is to extend the challenge as long as it remains a bottleneck to neuroscience, further enabling scientists to ask and answer many questions about the brain and its function that are currently beyond our reach.

3.2. The DIADEM Metric: Comparing Multiple Reconstructions of the Same Neuron

[With kind permission from Springer Science+Business Media: *Neuroinformatics*, volume 9, 2011, Todd A. Gillette, Kerry M. Brown, Giorgio A. Ascoli, © Springer Science+Business Media, LLC 2011]

Abstract

Digital reconstructions of neuronal morphology are used to study neuron function, development, and responses to various conditions. Although many measures exist to analyze differences between neurons, none is particularly suitable to compare the same arborizing structure over time (morphological change) or reconstructed by different people and/or software (morphological error). The metric introduced for the DIADEM (DIgital reconstruction of Axonal and DEndritic Morphology) Challenge quantifies the

similarity between two reconstructions of the same neuron by matching the locations of bifurcations and terminations as well as their topology between the two reconstructed arbors. The DIADEM metric was specifically designed to capture the most critical aspects in automating neuronal reconstructions, and can function in feedback loops during algorithm development. During the Challenge, the metric scored the automated reconstructions of best-performing algorithms against manually traced gold standards over a representative data set collection. The metric was compared with direct quality assessments by neuronal reconstruction experts and with clocked human tracing time saved by automation. The results indicate that relevant morphological features were properly quantified in spite of subjectivity in the underlying image data and varying research goals. The DIADEM metric is freely released open source (<http://diademchallenge.org>) as a flexible instrument to measure morphological error or change in high-throughput reconstruction projects.

Introduction

Neuronal morphology impacts network connectivity (Binzegger *et al.*, 2004; Lin and Masland, 2005; Stepanyants and Chklovskii, 2005) as well as electrophysiological function (Mainen and Sejnowski, 1996; Krichmar *et al.*, 2002; Van Ooyen *et al.*, 2002), including signaling propagation and integration (Vetter *et al.*, 2001; Schaefer *et al.*, 2003). The striking morphological diversity of both axonal and dendritic arbors reflects the complex form/function relationship in various neuronal types (Markram *et al.*, 2004; Goldberg *et al.*, 2006; Ascoli *et al.*, 2008). Morphological features are also examined to study the effects of environment (van Praag *et al.*, 2000), pathologies (Meyer-Luehmann

et al., 2008; Baloyannis, 2009), and development (Cline, 2001; Wong and Ghosh, 2002; Li *et al.*, 2005). Researchers have been interested in neuronal morphology for many years (Senft, 2011). In order to apprehend the factors that influence neuronal function beyond inspecting neurons under the microscope, rigorous analysis requires the reconstruction of morphology into measurable representations. Early efforts improved microscope capabilities and integration with computer control and measurement (Glaser and Vanderloos, 1965; Overdijk *et al.*, 1978), followed by attempts at automated two-dimensional automated neuronal reconstruction (Capowski, 1983; Mize, 1984). In recent history reconstruction has been largely accomplished by manual tracing with *camera lucida* (Sugihara *et al.*, 1996), through computer interfaces such as Neuromantic (<http://www.rdg.ac.uk/neuromantic>), Neurolucida (MBF Bioscience; Glaser and Glaser, 1990), and ImageJ plugins (Brown *et al.*, 2005) or even more recently using semi-automated tools like Autoneuron (MBF Bioscience), NeuronStudio (<http://research.mssm.edu/cnic>), V3D (Peng *et al.*, 2010a), and the TREES toolbox (Cuntz *et al.*, 2011).

In order to achieve the production scale necessary to answer the next generation of neuroscience questions, fully automated algorithms are needed along with adequate analysis methods. The development of a mammalian species connectome (i.e. a complete connection diagram of all neurons or neuron types) will require reconstructing millions to billions of neurons (Haug, 1987; Kasthuri and Lichtman, 2010). Even local or sparse connectivity studies may involve reconstruction of thousands to millions of neurons from multiple classes. In some cases, the same neurons will need to be traced at multiple time

steps, such as in developmental, plasticity, and genetic or molecular manipulation studies. Automatically traced reconstructions must meet appropriate quality levels for research. This raises the question of how to measure reconstruction error, and in turn to assess the underlying algorithm or process (e.g. a trainee learning to reconstruct). Any measure of quality is designed with particular goals in mind, which may differ among both algorithm developers and data users.

Accurate branch connectivity (i.e. topology, see **Figure 3.1a**) is particularly important for dendritic signal propagation and integration. Topology, along with trace centerlines (**Figure 3.1b**), may also be relevant for studying growth or pathology processes (Hao and Shreiber, 2007; Meyer-Luehmann *et al.*, 2008) and wiring principles (Stepanyants *et al.*, 2004; Marks and Burke, 2007). For connectomics, higher branch orders and terminal regions (**Figure 3.1c**) might hold greater relevance, since the precise path taken is less important than the actual synaptic locations in the terminal regions. However, for images with multiple overlapping neurites, particularly when there are overlapping branches that cross within the Z-resolution limit, the low branch order region would retain importance since a mistaken branch choice could lead to incorrect connectivity for an entire subtree.

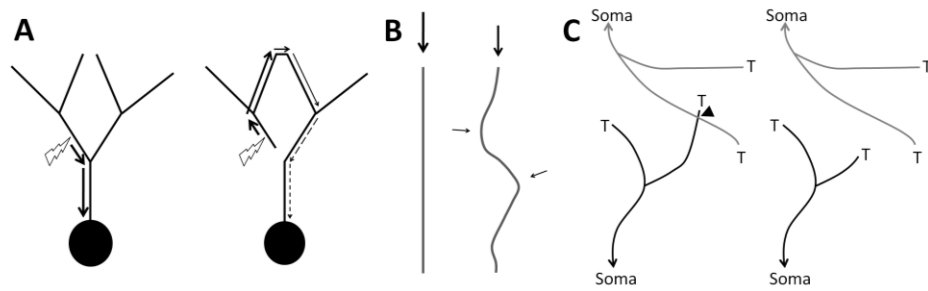


Figure 3.1. Effects of morphological features. **A.** Connecting the same dendritic points in different topologies has a large impact on charge transfer from synapse to soma, as shown by the varying responses (arrow size) to a stimulus in an identical location (*lightning bolt*). **B.** Branch centerlines determine path length and tortuosity, affecting signal propagation and providing clues about the underlying developmental processes. In terms of extrinsic determinants of branch growth, the left trace suggests a strong unidirectional signal (*arrow*) while the right trace reveals contributions from other directions. **C.** Precise tracing of dendrites and axons is important for determining potential connectivity. The left traces show a dendrite (*black*) and axon (*gray*) overlapping (*arrowhead*), while the right traces have no potential synapse due to the failure of the dendrite to extend all the way to the axon. Terminal nodes are all marked with a 'T'.

Scientists planning large-scale projects are crucially interested in the amount of their time an algorithm will save in producing quality reconstructions. While future algorithms may be even better and more consistent than humans at reconstructing neurons, the current state of the art requires checking and editing automatically produced reconstructions (Luisi *et al.*, 2011; Peng *et al.*, 2011). Thus, different scientists might disagree as to the most suitable criteria to decide on whether an algorithm suits their needs.

Human subjectivity in the tracing process presents a challenge to measuring reconstruction quality, particularly given the challenges associated with staining and imaging (Jaeger, 2001). Resolution limits of current imaging methods make it sometimes impossible to ascertain correct connectivity. Histological limitations such as tissue shrinkage, physical sectioning of trees, and labeling gaps and artifacts provide further demands on reconstruction methods (Brown *et al.*, 2011), and are important to consider

when analyzing morphological data (Kaspirzhny *et al.*, 2002). Methodological optimization and standardization of procedures, guided by collaboration between biologists and computer scientists, might reduce or eliminate such uncertainty by producing cleaner data. Until then, the metric design can reduce the effect of subjectivity by decreasing the weight of tree nodes (i.e. bifurcations and terminations) in more subjective zones, such as thin branches or distal locations, relative to the nodes in less subjective zones.

Many measures are commonly employed to analyze neuronal morphology, and new metrics continue to be introduced to capture specific features of dendritic and axonal arbors (e.g. Brown *et al.*, 2008). However, most studies tend to compare groups of neurons under differing experimental conditions. The typical design tests whether a sample of neurons from treated animals significantly differs from a sample (of the same cell type) from controls in terms of one or more morphometrics (e.g. number of branches or total length). These classical morphological studies are complementary to the problem of comparing individual instances of the same neuron. The question of how to measure the quality of a reconstruction relative to a gold standard is instead similar to the issue of quantifying changes in neuronal morphology over time or in response to some stimulus or treatment. In these cases, a metric that registers the tree structures and compares them node by node or branch by branch can provide greater detail and analytical power. Provided changes in morphological features are not so drastic as to change the neurite fundamentally and prevent proper registration, the detected changes could be attributed to particular segments of the neurite that appear, disappear, or change in some manner.

This paper presents a metric originally designed to measure reconstruction quality for the DIADEM Challenge (DIgital reconstruction of Axonal and DEndritic Morphology). The Challenge aimed to encourage development of automated algorithms with the goal of reducing human interaction time while still generating high quality data. The term “metric” is not used here in the strict mathematical sense, as the DIADEM metric produces a similarity value (on the scale of 0 to 1) rather than a distance value. Moreover, it is asymmetric (i.e. the score may depend on which reconstruction is considered the gold standard) and does not satisfy the triangle inequality. Although certain features in the original version of the metric were specialized for the competition, the current release (<http://diademchallenge.org>) expands the set of options and includes the source code for optimal suitability to a broader range of applications. After briefly reviewing prior efforts, we describe the DIADEM metric and examine its relationship with other components of the reconstruction quality judging process.

Previous Relevant Work

Although no metric exists specifically to measure change or error in neuronal morphology, several approaches have been developed to quantify features and reconstruction quality of tree-like structures for related purposes. One class of methods registers the same regions of interest and embedded trees in multiple medical images to each other given changes over time. These methods are commonly applied to human airways and vasculature, which are subject to elastic deformation due to effects such as breathing, blood flow, and surgery (Metzen et al., 2009). Registration is accomplished by matching nodes and/or branches in the arborization (Charnoz et al., 2005; Tschirren et al.,

2005; Drechsler et al., 2010) or by aligning path centerlines (Bülow et al., 2006). These approaches ignore spatial position since they account for tissue deformation (**Figure 3.2**). While these methods are broadly applicable to the goal of measuring morphological error and change, deformation is not an issue when comparing multiple reconstructions of the same neuron (see however Canty and De Paola, 2011). For this application, therefore, there is little reason not to incorporate relevant positional information in the desired metric.

The tree edit distance (TED) counts the number of nodes (both internal and terminal) that must be added or removed from one tree in order to create the second tree (Zhang, 1996). This pairwise metric has been applied to neuron classification (Heumann and Wittum, 2009) and to investigate stereotypy of neuromuscular fiber connectomes (Lu et al., 2009). However, TED is not spatially specific, while node locations are important in detecting morphological error or change. For example, when measuring quality, the double mistake of an extra branch and a missing branch extending from the same parent branch would be ignored by TED (see also Gillette and Grefenstette, 2009). The analogous situation for morphological plasticity would be the growth of one branch simultaneous with the retraction of another (i.e. within the same time frame, see **Figure 3.2**). Using an edit cost function (i.e., counting changes required to create perfect similarity between trees) for length differences would only capture part of the difference. A registration method accounting for position will avoid such a problem.

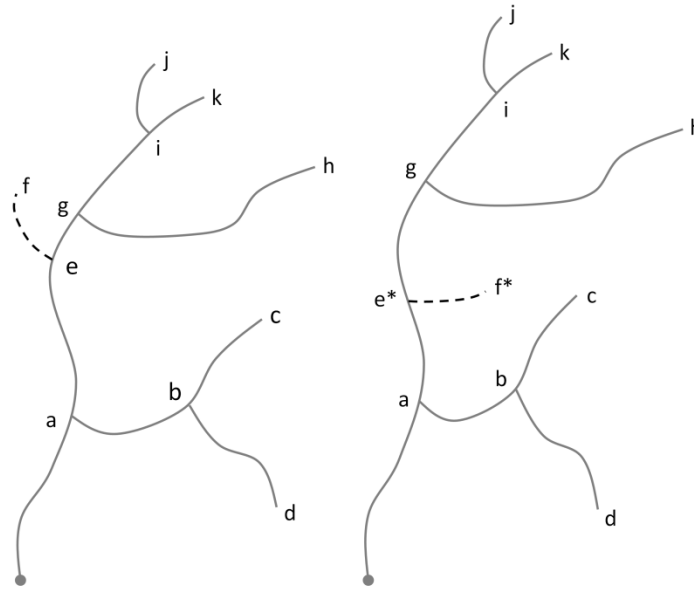


Figure 3.2. Tree registration and tree edit distance (TED). Methods that register nodes on deformed trees will match the dashed branches (nodes e and f with e* and f*), since topologically they are equivalent. Methods that include centerline or path matching will not match those branches. TED matches trees by their nodes and therefore will produce a distance score of 0 (i.e. a similarity score of 1). If TED distinguished the two branches, the distance score would be 4, i.e. one for each node added and removed (e, f, e*, f*).

Metrics developed to assess or validate algorithmic reconstructions in earlier automation efforts are also clearly of interest. The choice of validation strategy depends on the scientific aim or the perceived challenge. Previous selections included branch- and tree-level morphometrics (i.e. diameter, length, surface area), centerline accuracy, proportion of a gold standard reconstructed, and simulated electrophysiological activity (Wearne *et al.*, 2005; Tyrrell *et al.*, 2007; Losavio *et al.*, 2008; Rodriguez *et al.*, 2009; Peng *et al.*, 2010b). Simulations may be ideal to explore the impact of reconstruction quality on specific neuron functions. However, they are time-consuming and make limiting assumptions in terms of biophysical properties and stimulation protocols. Standard morphometrics suffer from the same issue described for TED, whereby a

parameter such as overall surface area could stay constant while different parts of the tree are changing in opposite ways. Distributions of those metrics may be more informative, but complete specification of all changes throughout the neuron requires registration and analysis of all branches.

Description of the DIADEM Metric

In comparing two arbors, the DIADEM metric registers and scores each corresponding node and its parent branch individually on the basis of connectivity with the local region of the tree. A rigid registration of nodes is used in order to make use of positional information, which is intuitively sensible when dealing with two reconstructions of the same underlying image data. In the case of comparing the same neurite at different times, the amount of change would have to be small enough that an identifiable core object is shared between time points. Adjustments for tissue deformation due to slice preparation or microscope stage drift would be necessary preprocessing steps. Emphasis on topologically relevant features over diameter and centerline geometry is justified in part by the influence of branch topology on signal processing and its dependence on growth mechanisms (van Pelt et al., 1992; Ascoli, 2002; Koene et al., 2009). Though vital to understanding neuron function, diameter evaluation is highly subjective at resolutions used for full neuronal arbor reconstruction (Jaeger, 2001; Scorcioni *et al.*, 2004), and was not considered in the DIADEM competition. Centerlines were deemed important in terms of reproducing distances along the neurite path. This aspect was incorporated in the metric without the need for continuous scoring of

centerline quality. The metric input is specified in SWC format¹, which represents trees by individual points in space with connections denoted by unique numerical identities. Below we explain the computational essence of the DIADEM metric. More extensive implementation details are provided in the online documentation and through distribution of the source code (<http://diademchallenge.org/metric.html>).

Each node in the “gold standard” arborization (e.g. the manual reconstruction) is taken as the ground truth and is registered to a node in the “test” arborization (e.g. the automated reconstruction). The process begins with the most proximal gold standard bifurcation, proceeds to the first bifurcation’s children, then to their children, and so on until registration has been attempted for all gold standard nodes. Registration requires the test nodes to be located spatially within a cylindrical threshold region around the gold standard node, as determined by the chosen data set or by user discretion. This represents the region in which one might reasonably create a bifurcation or termination in the trace given subjectivity and resolution in the XY plane and Z axis (**Figure 3.3a,b**). As either gold standard or test node could be at an edge of a bifurcation or termination, the radius of the threshold region was chosen to be the diameter of one of the larger bifurcations within each of the data sets used in the DIADEM competition in order to accommodate the more extreme (but still reasonable) cases. Image resolution, staining contrast, and depth of focus all influenced threshold choices such that sufficiently low resolution or contrast could result in larger thresholds if the neurite edges were less clear in the XY plane or Z-axis. Registered nodes are expected to have matching paths to matching

¹ <http://www.neuronland.org/NLMorphologyConverter/MorphologyFormats/SWC/Spec.html>

ancestor nodes (i.e. bifurcation nodes between the nodes of interest and the tree roots). Ancestor node matches are found by traversing toward the tree root from the gold standard node and from the potentially matching test node until the current ancestors of either tree are within the Euclidean distance threshold of each other (**Figure 3.3c**). The traversal takes place on one tree at a time, on the tree with the shortest path from the current ancestor back to its respective node of interest. Once an ancestor match is found, the path test is performed and the potential match is either confirmed or discarded.

To determine whether the gold standard and test paths match, an acceptable test path must have a sufficiently small error along a given component (i.e. on the XY plane or the Z axis) relative to the full three-dimensional path length. This “path length error” is calculated as the difference between the gold standard and candidate test path length in a given component, divided by the full gold standard path length. This method is one source of asymmetry in the metric. Another is the method for adjusting the test path length to account for differences in position of the ancestor and descendant from their respective gold standard matches (described in greater detail on the DIADEM Challenge website). Acceptable path length error is given by a second pair of thresholds, which for the DIADEM data sets were determined based on the straightest possible path and the longest reasonable path that stayed within the bounds of the neurite. Users may set their own path length error thresholds. A very large threshold might be chosen if path length is anticipated to be unreliable and unimportant in terms of data usage. If multiple test nodes are suitable for registration to a given gold standard node based on the criteria above, the one with a descendant and path to the descendant that matches a descendant (and path) of

the gold standard node is registered. If there are still multiple candidates or no candidates exhibit matching descendant connectivity, the test node spatially closest is registered.

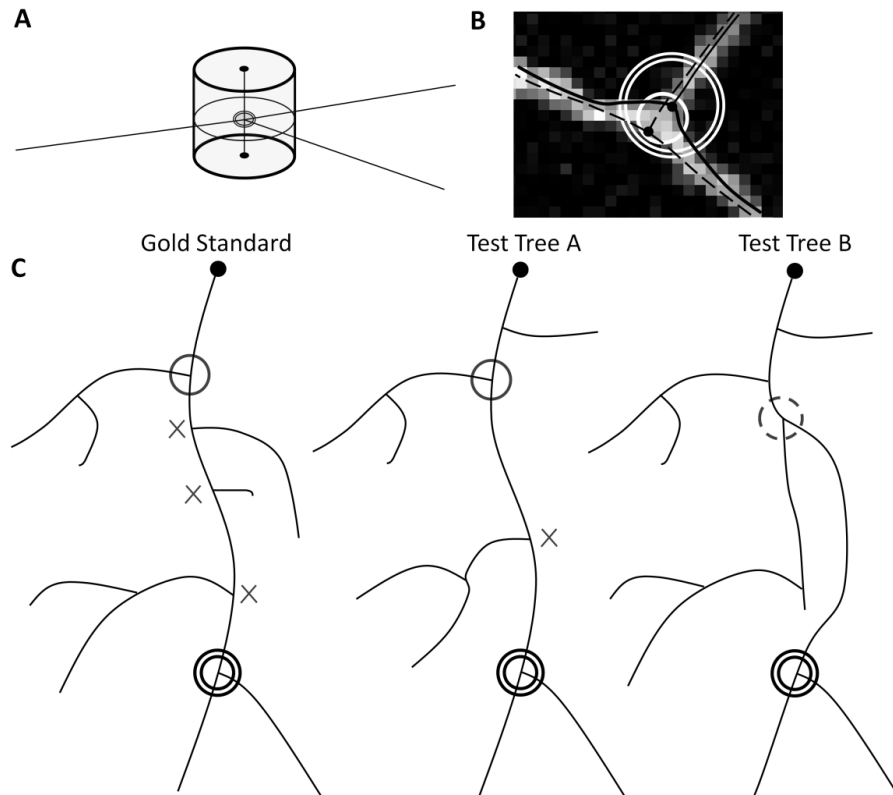


Figure 3.3. The DIADEM metric. **A.** The filled central circle intersected by the branches is the bifurcation node. The cylinder represents the 3D threshold region, within which potential matches may be located. The cylinder is generated from expanding the XY threshold circle to the bounds of the Z threshold. **B.** A trace could validly bifurcate anywhere within the specified region (single white ring). Here the gold standard (solid trace) bifurcates at the edge of the region. To account for the lack of information on the location of the neurite centerline, the threshold region (double white ring) has a radius twice that of the valid region. Thus a test trace (dashed) that bifurcates at the opposite edge will still be matched. **C.** Establishing branch connectivity with an ancestor. The target node of the gold standard reconstruction (double ring) has a candidate test node in both test trees to the right. The \times symbols are the ancestors of the target node that have no match in Test Tree A, and the ancestors of the potential match in Test Tree A that have no match in the gold standard. However, the black ring ancestor matches spatially and in terms of path to target, resulting in a hit for Test Tree A. Test Tree B has an ancestor in the right location, but the length of its path to the target node is incorrect (dashed ring), resulting in a miss for Test Tree B.

Once registration has been attempted for all gold standard nodes, those nodes that fail to be registered to a test node are each considered a second time to determine whether the test tree at least captures the path through the location of the gold standard node. This situation can occur when there is a bifurcation in the gold standard but only one of the two child branches is traced in the test reconstruction resulting in a single path with no bifurcation node to use for registration. Without accounting for these cases, test reconstructions would be counted as having missed the path leading up to the gold standard bifurcation node even though it was traced (**Figure 3.4**). The matching path is found by first traversing the gold standard reconstruction toward the root. When an ancestor that has been registered to a test node is found, the gold standard tree is then traversed from the target gold standard node toward its terminal ends. If no test descendent nodes match the path between this matched ancestor and the two gold standard descendants, paths to nodes further downstream are checked for matches. The number of descendant paths followed depends on the number of downstream bifurcations, but the traversal of any given path ends when a descendant is encountered that is registered to a test node. The same criteria for node matching described above applies here for matching descendants, including having similar path lengths to the ancestor of the gold standard node in question. The process is continued until a path match is found, or until all potential descendant paths are exhausted. If a matching path is found, the test tree is scored as having a match to the gold standard node since the test path is deemed to have captured the same path up to the node location on the gold

standard node. The absence of the second branch is penalized in the initial registration process by the lack of a match of its terminal node (or subtree).

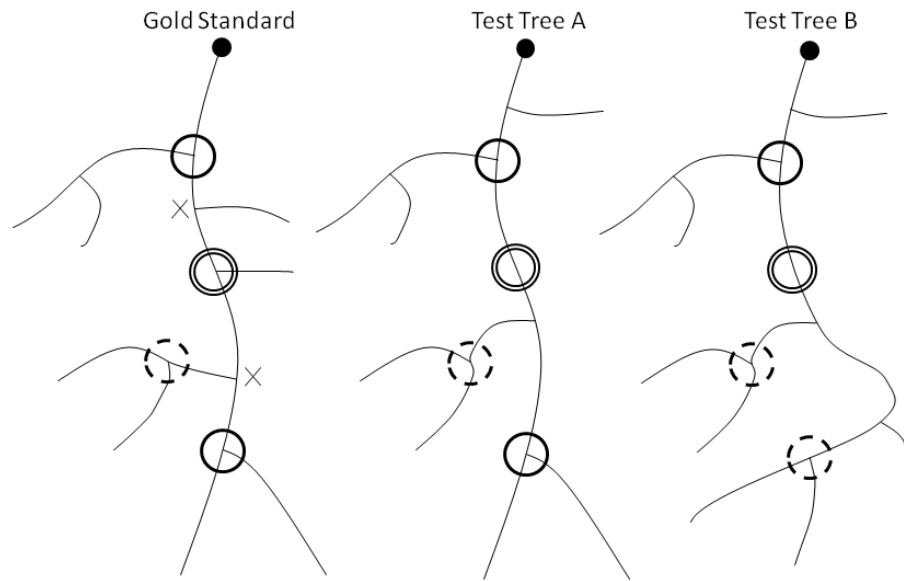


Figure 3.4. Matching paths through unregistered nodes. Symbols on the gold standard correspond with symbols on Test Tree A, which has a matching path (thick path) for the unregistered gold standard node of interest (double ring). Test Tree B results in a miss for the gold standard node and path. The × symbols mark an ancestor and two descendants without matches in either test tree. Both test trees do contain a matching ancestor (ring above target). Several descendants have potential matches (dashed rings and their descendants), but fail to produce matching paths to the ancestor. The position of the second descendant (solid single ring below target) and its path to the ancestor do match in Test Tree A. No descendants in Test Tree B match by path, and thus Test Tree B does not contain a matching path for the gold standard target node.

Several alternative options provide greater flexibility depending on the intended use of the metric. Nodes can be weighted differentially, for instance by degree (i.e. number of terminal nodes in the sub-tree) in order to capture topologically important bifurcations, or uniformly if terminal nodes are as relevant to the user. In the DIADEM competition, nodes were weighted by degree primarily due to the larger functional impact

of larger sub-trees. In several of the DIADEM gold standards, tracing subjectivity was also particularly high in terminal regions, with more frequent branching and neurite overlap (Brown *et al.*, 2011). Thus, this weighting scheme also reduced the contribution of the more subjective regions to the score. Another choice is whether to tally as misses the excess nodes in the test reconstruction that are not in the gold standard. Rules are in place to prevent non-matched test nodes that appear to be reasonable attempts (i.e. the bifurcation or termination is close to a gold standard node) from counting as excess nodes (**Figure 3.5**). The miss in the gold standard in these cases is considered to be a sufficient penalty to the score. If used for measuring changes in neurite morphology, a suitable alternative could be to consider all unmatched test nodes as new branches by counting them all as excess nodes rather than using any criteria to excuse them. Lastly, terminal test nodes that terminate in the wrong location can be scored as matches if their path is approximately correct. While not part of the default mode, this option was used in order to prevent penalizing automated algorithms that traced rosette structures in the neuromuscular projection fiber data set (**Figure 3.6**). Further potential uses of this option are discussed later. In applying the DIADEM metric, the weighting scheme and other parameters must be determined based on a combination of goals, including minimizing subjectivity and maximizing functionally relevant features.

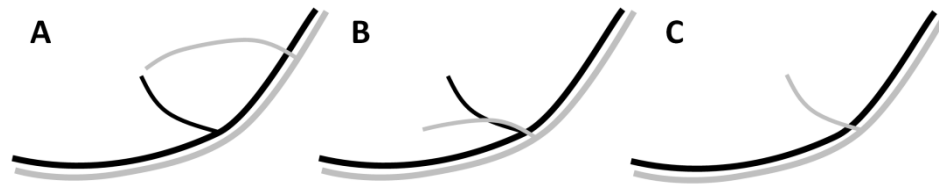


Figure 3.5. Determining excess nodes. **A.** The terminal node of the gold standard (black) is a miss due to its incorrect path. The gray test termination is not an excess node because there is a gold standard node within threshold distance. Since the termination does not count as an excess node, the parent bifurcation is not counted against the score either. **B.** The gold standard terminal node is a miss due to the incorrect position of the test terminal (gray). However, the test termination is not an excess node since its parent bifurcation is matched. **C.** The gray test termination has no basis in the gold standard reconstruction and so it is an excess node.

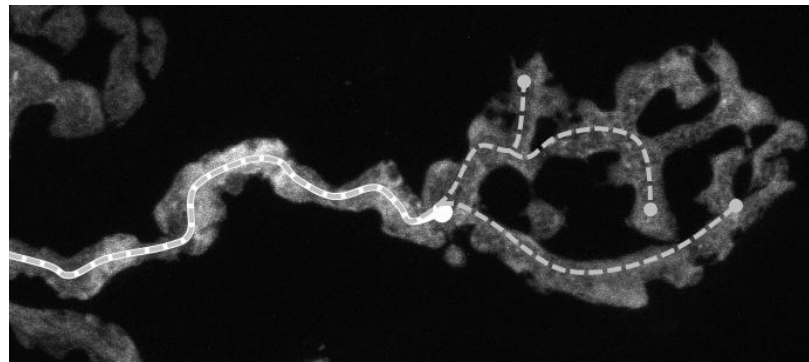


Figure 3.6. Neuromuscular projection axon rosette. The gold standard (*solid white trace*, which overlaps with the *dashed gray line*) tracks the path of the axon until it ends at the beginning of the rosette structure. The test reconstruction (*dashed gray*) follows the same path yet continues into the rosette, and will still be successfully matched. This approach could also be used for measuring the growth of a neurite captured across multiple reconstructions in a time series.

Human Expert Qualitative Scores and Interaction Time

In alternative or addition to an objective metric, the quality of a reconstruction can be evaluated directly by a domain expert. After the Qualifier Round of the DIADEM Challenge (Brown et al., 2011), several experienced neuroscientists judged the quality of the submitted reconstructions. All judges were principal investigators of laboratories that had produced several peer-reviewed publications using neuronal reconstructions.

Although no strict constraints were mandated as to how to score submissions, guidelines were provided in terms of the possible reconstruction errors that should be looked for. These included missing a large sub-tree, missing small branches, tracing non-existing or unconnected trees, tracing extra small branches, incorrectly positioning bifurcations and terminations, incorrectly connecting branches and trees, incorrectly tracing a segment path, and tracing a segment in the wrong direction. The judges determined for themselves how to weight each type of error. Their scores ranged from 0 to 10, but were rescaled 0 to 1 for comparison purposes with other metrics. In the analysis these scores are referred to as “qualitative scores”. “Average qualitative score” will refer to the average of scores from multiple judges for one automated reconstruction.

Reconstruction algorithms cannot yet consistently produce reconstructions of gold standard quality. Nevertheless, their output may be of sufficient quality such that editing it could take considerably less time than reconstructing the same neuron from scratch. Thus, the effectiveness of an automated reconstruction algorithm can also be measured by an expert determination of the amount of time saved compared to a manual process (Chklovskii et al., 2010). A person highly experienced in reconstructing a neuron from a particular preparation can produce an accurate estimate by clocking the interaction time taken to reconstruct a neurite sample from scratch. Then, the same operator clocks the time necessary to edit the reconstruction of an equivalent sample produced by the automated algorithm until it is of the same quality as that generated manually. The resulting score, hereto referred to as the “interaction time score” or simply “interaction score”, would be specific to the type of data and may contain some subjectivity reflecting

the individual's strengths and weaknesses in terms of speed in reconstructing or editing specific morphological features. However, the score directly addresses the scientist's main concern by quantifying the time saved by the evaluated algorithm.

Submissions of the DIADEM Qualifier Round were scored according to this criterion in addition to applying the DIADEM metric and the expert qualitative assessment. In particular, the owners of the data sets used in the competition scored the entries by determining how long it would take to edit them up to gold standard quality without simultaneously viewing the gold standard. Experience with a particular image stack could lead to a decrease in editing time for a data owner independently of reconstruction quality. In order to minimize the relative improvement between reconstructions from different teams of the same neuron, the data owners were instructed to start by scoring the manual reconstruction. Data set owners also shuffled the order in which team reconstructions were edited. While this second factor reduced bias toward any team, it would not eliminate the experience effect on scores for the analysis comparing the interaction time and the metric. Data set owners only scored reconstructions of their own neurons, and in many cases only scored a subset of those. For any submission that was edited, the judge either reconstructed the section or estimated the amount of time it would take to do so, but in either case was consistent in their practice. The corresponding score for a given reconstruction was one minus the ratio between the amount of time needed to edit the reconstruction and the total time taken to manually reconstruct the same structure. Thus a score of zero would correspond to no

time savings, and a score of one would correspond to a perfect automated reconstruction requiring no editing time.

Utility of the DIADEM Metric

In order to determine whether the DIADEM metric is useful for evaluating reconstruction quality, it is important to compare it to expert judgments. Indeed, comparison between the various scoring methods showed strong correlations, suggesting that the metric may be viable as a machine-derived surrogate to human judging (**Figure 3.7**). Specifically, the correlation across reconstructions between the metric and the qualitative score averaged from four different judges was $R = 0.66$ ($p < 10^{-5}$). Correlation between the metric and interaction time scores was $R = 0.77$ ($p < 10^{-6}$) when controlling for judge and data set. All correlations with interaction time scores are determined using a linear model of interaction time score as the dependent variable, and the correlate (in this case the metric score) and data set as the independent variables, without any interaction effects. The R value is derived from the resulting adjusted R^2 of the model. This normalization step is necessary because, while each qualitative judge scored every data set, each data owner only clocked the interaction time for their own data set. Thus interaction times could be affected both by the judge's methods and data specific features, as confirmed by the statistically non-significant correlation of $R = 0.09$ ($p = 0.6$) in the absence of control.

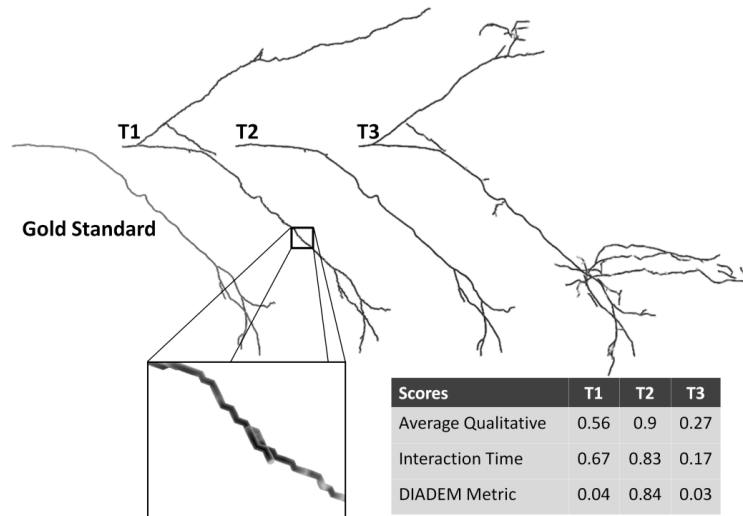


Figure 3.7. Scoring method comparison. The gold standard and automated reconstructions (T1-T3) of a hippocampal CA3 interneuron dendrite submitted in the Qualifier Round of the DIADEM Challenge. The qualitative score, interaction time, and DIADEM metric for each reconstruction exemplify how various errors can impact different scoring methods. On this small tree, T1 received a low DIADEM metric score due to unsmoothed paths which inflated path length (*zoom box*) causing path length error to surpass the threshold. A smoothed path or more lenient path length threshold would result in a higher score more consistent with human expert judgment which ignores the effect of the jagged path. In T3, too many excess nodes resulted in low scores for all measures.

If the data set effect is in fact a judge effect, then the metric can better serve as a comparison than interaction time scores produced by different judges (at different labs). However, variability in interaction time scores could be due to tree size, which varies substantially among data sets (**Table 3.1**). If tree size explained the score variability, a model of metric and interaction time including tree size should show a strong tree size effect and a correlation similar to that observed in the model that included data set as a parameter. Instead, such a model produced a poorer correlation. Moreover, a model including both data set and tree size showed no significant effect of tree size. Therefore, while tree size may have some effect on scores, it is clearly not responsible for variation in interaction times between data sets. Further evidence in this regard comes from the

variability in qualitative judge scores. The four qualitative scores for each of the selected reconstruction submissions had an average range of 0.3 and an average standard deviation of 0.14. While experts may each be consistent in their judgments, these results suggest that an objective measure such as the DIADEM metric will better serve for comparing reconstruction qualities across studies.

Table 3.1. Scores across data sets. Values are means \pm standard deviations. Tree size refers to the number of branches of the gold standard reconstructions. The Climbing Fiber data set had only one gold standard reconstruction, thus it has no standard deviation. The “Uniform” column represents values from the DIADEM metric with nodes weighted uniformly.

Data set	Tree Size	Qualitative	Interaction	DIADEM	Uniform
Climbing Fiber (n=5)	199	0.46 ± 0.27	0.11 ± 0.18	0.37 ± 0.28	0.22 ± 0.16
Hippocampal CA3 (n=14)	4 ± 2.2	0.58 ± 0.26	0.66 ± 0.16	0.39 ± 0.32	0.44 ± 0.31
Neocortical (n=13)	8.2 ± 7.5	0.4 ± 0.17	0.41 ± 0.17	0.28 ± 0.3	0.25 ± 0.27
Olfactory Projection (n=8)	26 ± 9.6	0.8 ± 0.06	0.3 ± 0.15	0.78 ± 0.1	0.58 ± 0.09

The correlation between the DIADEM metric and either of the other two methods (qualitative evaluation and interaction time) is nearly as strong as the correlation between the two human scores, indicating that the metric might capture fundamental features shared by expert judgment. Again controlling for data set, the correlation between qualitative scores and interaction time was $R = 0.83$ ($p < 10^{-8}$). A linear model controlling for data set also confirms the data set effect in the interaction scores. For example, the hippocampal CA3 interneuron ($p < 10^{-7}$) and neocortical layer 6 axon ($p < 10^{-4}$) data sets had significant effects when the cerebellar climbing fiber data set is taken as the base

model (**Figure 3.8**). This was the same pattern observed in the model of DIADEM metric and interaction time scores. While each measure of reconstruction quality (relative to a gold standard) is unique, the analysis suggests that the measures all are strongly related to each other and that the DIADEM metric can serve as a surrogate for expert qualitative and interaction time judgment (see **Table 3.2** for summary of correlations).

Table 3.2. Summary of correlations. All correlations are highly significant ($p < 0.001$). Values mirrored over the diagonal for ease of comparison².

	DIADEM Metric	Qualitative	Interaction Time	TED
DIADEM Metric		0.66	0.77	0.52
Qualitative	0.66		0.83	0.54
Interaction Time	0.77	0.83		0.75
TED	0.52	0.54	0.75	

The effects of specific scoring method differences are more difficult to tease apart. Some of the remaining variance could in principle be attributed to the factors that are differentially weighted in each scoring system. For instance, while the DIADEM metric weights nodes by terminal degree, the interaction time score does not necessarily differentiate nodes by sub-tree size. However, a variant of the metric with uniform weights produced no better correlation with interaction time scores ($R = 0.75$) or qualitative scores ($R = 0.63$). As an alternative explanation, judges might not necessarily

²Values are Pearson correlation coefficients, except for those associated with Interaction Time (gray background), which are each calculated as the square root of the adjusted R^2 of the linear model. The models take Interaction Time as a function of data set and a specific scoring method.

mark a branch as incorrect just because it terminates too early or too late, thus differing from the DIADEM metric (**Figure 3.8b**). Another source of variance that could not be controlled for was the increased experience of interaction time judges over a set of reconstructions of the same underlying neurite. Shuffling the order across team reconstructions alleviates bias for or against teams, but it is still possible that two reconstructions of the same neurite and of equal quality could receive different interaction time scores. Despite these sources of variance and other unexplained sources, the correlation results suggest that the DIADEM metric can serve as a complementary judgment of features that are more difficult to visualize, particularly branch topology (**Figure 3.9**). Different connectivities within two complex trees can have a profound effect on signal propagation and/or integration, even if spatially similar and thus easily missed by the human eye.

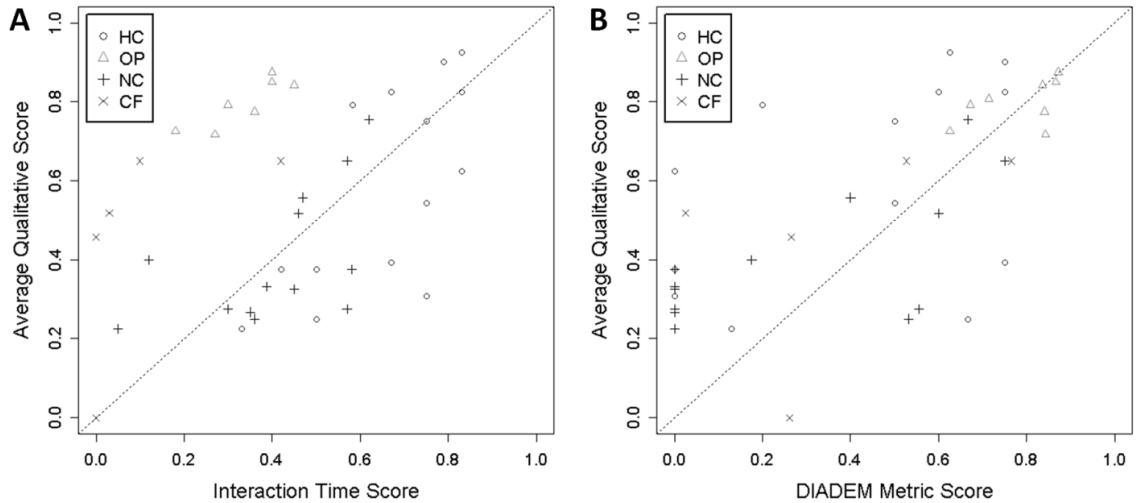


Figure 3.8. Relationships between scoring methods. Data sets are identified with different symbols and abbreviations (HC: hippocampal CA3 interneuron; OP: olfactory projection fibers; NC: neocortical layer 6

axons; CF: cerebellar climbing fibers). All scores were normalized to a scale of 0 (incorrect) to 1 (perfect). **A.** Separation of scores by data set clearly shows the effect of data set on interaction times. **B.** The lower DIADEM metric scores (along the left edge) are responsible for much of the lack of correlation. These scores are largely due to terminal branches that track the gold standard well, but terminate in the wrong location. The effect is much stronger in the data sets containing trees with fewer branches (HC and NC).

The importance of positional information is demonstrated by comparing the DIADEM metric to the tree edit distance (TED) described above (Heumann and Wittum, 2009). In order to compare the two measures, tree edit distances values were normalized by the gold standard tree size and inverted in order to generate comparable similarity values. TED correlations with qualitative scores ($R = 0.54$) and with interaction times ($R = 0.75$) were lower than those of the DIADEM metric. The correlation between TED and the DIADEM metric improved from 0.52 to 0.68 when considering only gold standard reconstructions with degree greater than 3 (more than 5 branches). This change, and much of the lower correlation, could be due to the fact that two trees could have the exact same topology yet differ in terms of branch locations. This is a substantial risk particularly for small trees with few distinct topologies. The higher DIADEM scores may reflect the treatment of continuations as hits, where cost is calculated at the missing descendants, as TED tallies continuations as deleted nodes, and each extra branch as two added nodes. Recognizing continuations as hits is important as the node also represents its parent branch. Thus, the DIADEM metric is more appropriate than TED for the task of measuring reconstruction quality or morphological change of a neuron.

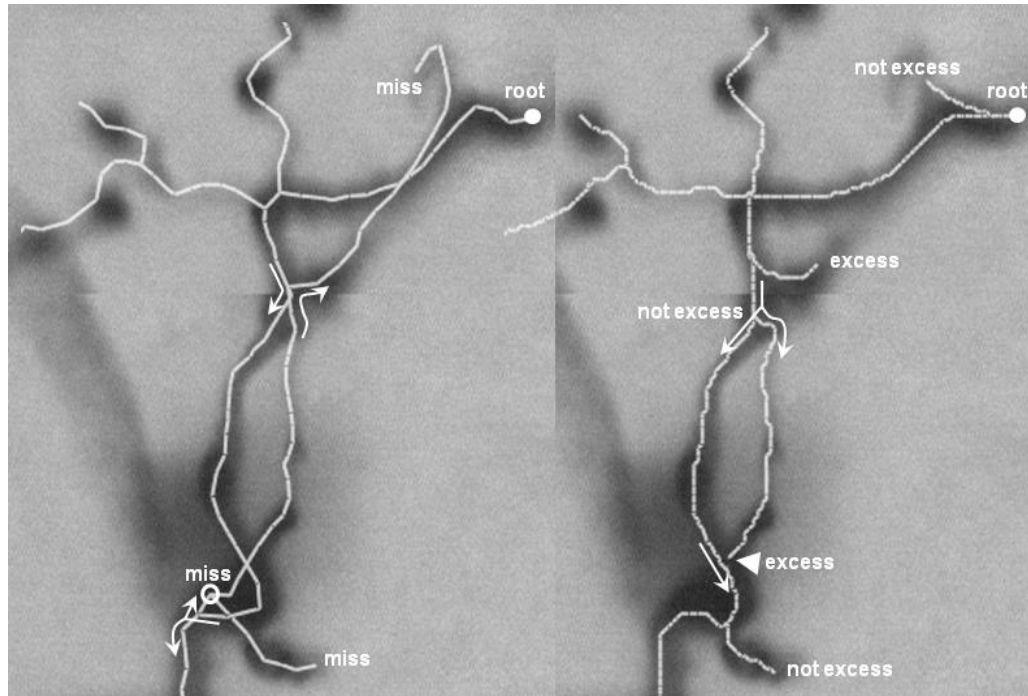


Figure 3.9. Tracing errors. Selected region of cerebellar climbing fiber single slice with gold standard overlay (*left*) and sample automated reconstruction overlay (*right*). While some errors in the automated reconstruction are more obvious, the mistake in the central region is more difficult to discern without close attention. The right side of the gold standard trace runs from a bifurcation near the bottom to the top, but the automated reconstruction creates a non-existent bifurcation which leads the trace downward and stops, yielding an excess terminal node (*arrowhead*). While the bottom bifurcation (bifurcating *white arrows*) is seen as a continuation, the error produces another missed bifurcation immediately above it (*hollow circle*) and then an additional two missed terminations.

Discussion

The DIADEM metric was designed to quantify the differences between two reconstructions of the same neuron. Its ability to provide useful scores of reconstruction quality was confirmed by strong correlations with expert qualitative judgments and measures of interaction time required to fix reconstruction errors. Thus, the DIADEM metric could greatly aid the development and tuning process of automated reconstruction algorithms. The metric also has particular qualities that would make it useful for

measuring morphological change of a neuron over time due to development, plasticity, or other experimental factors.

The DIADEM metric is released open source to encourage future implementation of further improvements. A more sophisticated path matching module would decrease the likelihood of false positives (incorrectly registering nodes) while creating an option to score a node based on its parent branch centerline quality. An average distance method used on vasculature centerlines (Schaap *et al.*, 2009) could be adapted to the DIADEM metric, given the known starting and ending locations of the path. Such an approach would match points along both reconstruction paths, then find the average distance between paths. Alternatively, a sequence alignment method using position and curvature (direction) could also increase path matching accuracy (Cardona *et al.*, 2010). This method would be particularly useful in cases of deformation where a larger position threshold is required and a simple average distance is insufficient.

While a direct comparison of the DIADEM metric with the basic TED showed the metric to better represent expert judgment, the potential of TED is worth consideration for use in a future improved metric. Heumann and Wittum (2009) tested a variety of node labels, evaluating matched nodes and assigning some cost to differences in matched node properties. A variant TED might be designed that uses node position as part of a node label, and using a nonlinear node edit distance function that evaluates to 0 within a threshold range, and to infinity beyond that range. This would at least handle the issue of node position, one of the primary characteristics of the DIADEM metric. Such a scheme would also allow inclusion of other branch features such as path length or surface area.

This would be even more useful in measuring morphological change. Enabling non-binary scoring for each node based on branch features would create options particularly suited to measuring differences in a neuron over time, such as terminal retraction or elongation. Nevertheless, these variations on the TED alone would still fall short of the DIADEM metric functionality. An advanced DIADEM inspired metric might incorporate those aspects of the TED into an algorithm that also weights nodes, handles continuations, and makes allowances for certain unmatched test nodes (i.e. excess node testing). Moreover, in order to capture qualitative judgment of terminal nodes that are too long or too short (or to properly measure growth or retraction of terminal nodes), terminal node matching would require specialized processing such as that applied in the neuromuscular projection fiber rosette case (**Figure 3.6**). Thus, although appropriate TED variations might approach the capabilities of the DIADEM metric, only in merging the two approaches would its benefits be fully realized.

Repeated gold standards are invaluable to account for subjectivity in evaluating reconstructions. A method for measuring the quality of algorithms to reconstruct arterial centerlines used a reference composed of several manual reconstructions produced by different people (Schaap *et al.*, 2009). While averaged centerlines (and their variances) can be used to evaluate paths between nodes, the topology of trees cannot be averaged since the paths between nodes either do or do not exist. However, if multiple gold standard reconstructions are available, a test reconstruction node and its path could be considered valid if matches were present in a representative sample of the gold standards. Overall, an automated algorithm should be deemed of “gold standard quality” if its

product falls within the variability displayed by the manual reconstructions generated by several independent experts. A more detailed analysis would be required for any data set to determine how many manual reconstructions are needed to capture the full range of subjectivity.

A metric that tests automated reconstructions can also serve in an iterative process of improving the algorithm or tuning its parameters. While developers may choose to use it as a separate step, the DIADEM metric can also be integrated into a development pipeline that automatically searches optimal parameters for a given data type. The metric can return missed nodes, nodes marked as continuations, and excess nodes. It could further be enabled to output other details such as centerline precision. Such information could aid identification of the most challenging features for the reconstruction algorithm, as well as tracking of the changes that produce improvements for specific node types.

In conclusion, assuming a manual reconstruction is available to use as the gold standard, the DIADEM metric captures a substantial proportion of expert qualitative judgment and editing time, and it can also detect topological errors easily missed by humans. Future implementations utilizing the ability of the metric to register topology and measure node-specific differences will enhance its potential to also quantify neuromorphological change in time.

Information Sharing Statement: The DIADEM metric executable is a cross-platform JAR file which can be run by the Java Virtual Machine on any computer or operating

system. The executable, code, documentation, and instructions for use are available at <http://diademchallenge.org>.

Acknowledgements

We are grateful to Dr. Karel Svoboda for early discussions on the development of the DIADEM metric. This work was supported in part by HHMI and NIH grant R01NS39600.

CHAPTER FOUR: TOPOLOGICAL CHARACTERIZATION OF NEURONAL ARBOR MORPHOLOGY VIA SEQUENCE REPRESENTATION. I. MOTIF ANALYSIS

[An edited version of this chapter is in press: Todd A. Gillette, Giorgio A. Ascoli: in BMC Bioinformatics (2015)]

Abstract

Background: The morphology of neurons offers many insights into developmental processes and signal processing. Numerous reports have focused on metrics at the level of individual branches or whole arbors; however, no studies have attempted to quantify repeated morphological patterns within neuronal trees. We introduce a novel sequential encoding of neurite branching suitable to explore topological patterns.

Results: Using all possible branching topologies for comparison we show that the relative abundance of short patterns of up to three bifurcations, together with overall tree size, effectively capture the local branching patterns of neurons. Dendrites and axons display broadly similar topological motifs (over-represented patterns) and anti-motifs (under-represented patterns), differing most in their proportions of bifurcations with one terminal branch and in select sub-sequences of three bifurcations. In addition, pyramidal apical dendrites reveal a distinct motif profile.

Conclusions: The quantitative characterization of topological motifs in neuronal arbors provides a thorough description of local features and detailed boundaries for growth mechanisms and hypothesized computational functions.

Introduction

Neuronal morphology is determined by a number of factors, including physical and biological constraints and requirements of axonal, dendritic, and network function. Branching topology is a complex feature of arbor morphology and is generally measured via one of several metrics: number of branches, maximum branch order (i.e. number of bifurcations between root and tip), partition asymmetry (Van Pelt *et al.*, 1992), and caulescence (i.e. prominence of a main path [Brown *et al.*, 2008]). While these metrics have proven useful in many studies, they do not necessarily capture the detailed branching patterns of neurons. We introduce a method for representing a neuronal tree as a sequence of characters, each encoding for select features of a branch. We tested this method on a large set of neuronal reconstructions from NeuroMorpho.Org (Ascoli *et al.*, 2007). Specifically, we analyzed the branching sequences for motifs to identify patterns (subsequences) representative of arbor types (axons, dendrites, and pyramidal apical dendrites).

Branch diameter and neurite length have long been known to impact passive and active electrical propagation (Rall, 1969; Zador *et al.*, 1995), and branching patterns can influence the order and timing of input signal integration and postsynaptic receipt (Jeffress, 1948; Van Ooyen *et al.*, 2002). Ultimately it is the combination of features, including non-morphological features such as channel composition and density, which

determine electrophysiological function. Moreover, arbor size and branching patterns reflect the distribution of synaptic (pre or post) targets of a neuron given metabolic and volumetric constraints (Cuntz *et al.*, 2007, 2012; Wen and Chklovskii, 2008). Beyond the functional focus, branching features result from particular growth processes and thus can be used to validate growth models driven by biophysics (Hillman, 1979; Tamori, 1993), molecular and signaling dynamics (Chklovskii, 2000; Hely *et al.*, 2001; Luczak, 2006; Torben-Nielsen and De Schutter, 2014), statistical relationships (Burke *et al.*, 1992; Donohue and Ascoli, 2005b, 2008; Samsonovich and Ascoli, 2005), or more abstract positional and branch-type rules (van Pelt and Verwer, 1983; Van Pelt and Verwer, 1986).

Until recently, the challenges of generating relevant biological data have limited cell type growth analyses to a single type or a few related cell types at a time. However, accelerating data production and curation of datasets from a growing number of laboratories by NeuroMorpho.Org have enabled large-scale analysis. Some studies have used these data to detect general principles of neuronal organization (Teeter and Stevens, 2011; Cuntz *et al.*, 2012). New imaging and semi- to fully-automated reconstruction processes (Donohue and Ascoli, 2011; Liu, 2011) are contributing to increased throughput, with ever larger datasets to be expected in the near future (He and Cline, 2011). For detecting relationships between cell types, this influx of data calls for new methods of analysis (Polavaram *et al.*, 2014).

Other fields have seen similarly dramatic growth in data, most notably genomics. Statisticians and computer scientists responded by creating a wide and still growing array

of techniques to sort through the data in a practical timeframe in search of significant relationships and findings. These techniques broadly include local and global alignment (Altschul *et al.*, 1990, 1997; Brudno *et al.*, 2003); as well as multiple sequence alignment (Thompson *et al.*, 1994), each with many specific algorithms designed for targeted sensitivity and/or efficiency. We leveraged the underlying bases of these techniques to analyze neuronal morphology by representing axonal and dendritic trees as sequences of branches encoded by their features. We began by sequentially encoding local topology based on a simple binary tree representation of neuronal structure. We then applied motif analysis (Milo *et al.*, 2002) to determine the defining topological patterns across arbor types from a broad range of species, neuron classes, and brain regions. Highly over- or under-expressed patterns constitute motifs and anti-motifs, respectively.

Methods

All code is available open source at <http://krasnow1.gmu.edu/cn3/NeuriteSequence/>, including Java implementation, R analysis scripts, and related documentation.

Neurite trees to sequences

The possible ways to encode a neurite as a sequence are numerous. As the first and simplest approach, we used local topology alone for the encoding of bifurcation nodes. Specifically, bifurcations are encoded on the basis of whether their child branches lead to bifurcations or terminations. Bifurcations in which both child branches themselves bifurcate are encoded with the letter ‘A’ (for “arborizing”). Bifurcations with one bifurcating child and one terminating child are encoded as ‘C’ as the tree “continues”

without adding a new subtree. Bifurcations with two terminating children are encoded as ‘T’ (**Figure 4.1a**). These definitions are equivalent to those used in vertex analysis (Berry and Flinn, 1984) with the A, C, and T bifurcation types referred to as tertiary, secondary, and primary nodes. Note that terminal branches, though not explicitly encoded, are fully accounted for in this method.

To string the character representations into a sequence, neuronal trees must be traversed. We use a traversal that encodes the root node of a tree and then recursively selects each subtree of that root in succession (i.e. prefix, depth-first traversal). One of the child subtrees is fully traversed first, followed by the second child subtree, and the letter encoding each encountered node during traversal is appended to the sequence (**Figure 4.1b**). This method optimizes the locality of representation, keeping nodes that are adjacent in the tree structure as close as possible in the sequence. Example sequences of the apical dendrite, basal dendrite, and axon of a pyramidal cell can be seen in **Figure 4.1c,d**. The order of subtree traversal is consistently determined by the number of bifurcations in each subtree. We investigated topological sequencing produced by always traversing the smaller subtree first (Smaller then Larger: *StL*) as well as those produced by always traversing the larger subtree first (Larger then Smaller: *LtS*). The *StL* representation has a greater locality, with adjacent tree nodes being nearer in the resulting sequence, than the *LtS* representation. All sequences and *k*-mers shown here are based on *StL* traversal, though the full analysis also includes unique *LtS* trees, as discussed in Section 2.2. In either *StL* or *LtS* traversal, when both subtrees of a node have the exact same number of bifurcations, the order is determined by the subtree topology, as

described by Harding (1971). Specifically, trees with higher partition asymmetry are treated as the larger subtree.

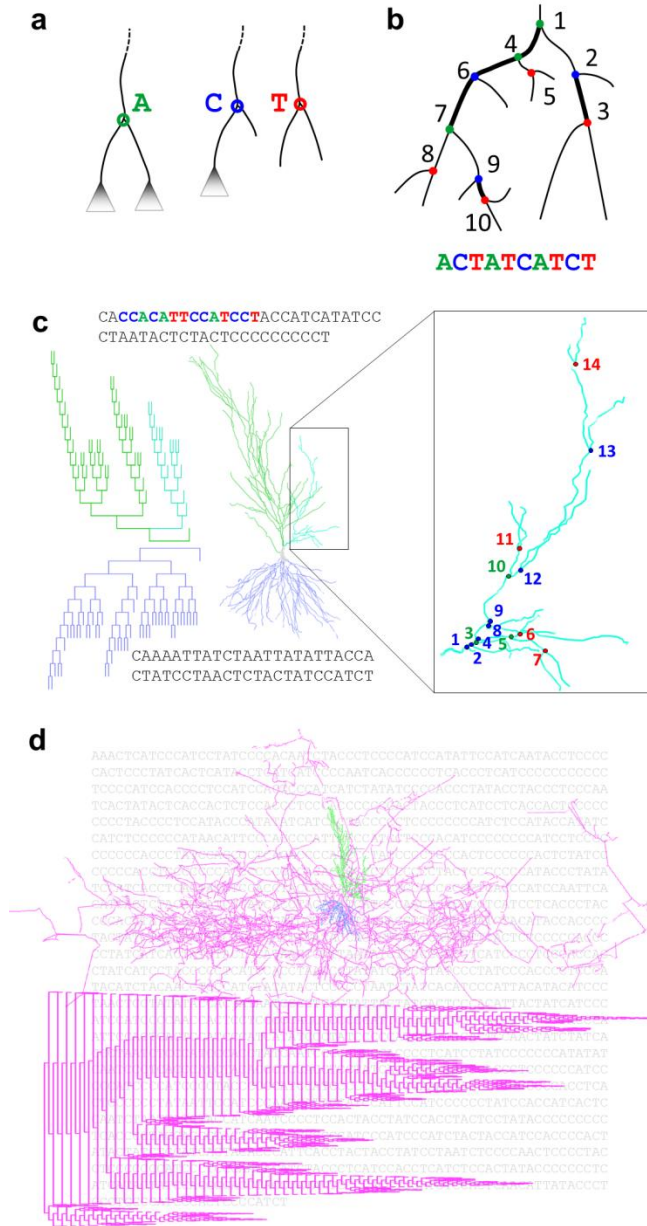


Figure 4.1. Converting morphologies to sequences. **a.** Bifurcation nodes are encoded as characters based on whether their child branches bifurcate or terminate. Arborization (**A**) nodes have two bifurcating children. Continuation (**C**) nodes have one bifurcating and one terminating child. Termination (**T**) nodes have two terminating children. **b.** Nodes are traversed depth-first starting from the smaller side which

optimally preserves locality. **c.** Hippocampal pyramidal cell apical (green) and basal (blue) dendrograms and morphologies are shown (NMO_00191 from [Turner et al., 1995]), with enlargement of a portion of the apical dendrite (right) and coloring in the sequence. Node types are colored and numbered by their order in the sequence starting with the first node in the subtree. **d.** The entire pyramidal cell morphology is shown (top), with dendrogram (bottom) and sequence representations (background) of the axonal arbor (magenta) (NMO_07897 from [Ropireddy et al., 2011]).

K-mers

To determine what patterns neurites exhibit among all tree shapes, a motif analysis was carried out for bifurcation subsequences of (increasing) length k , termed k -mers. Besides the three monomers A, C, and T, there are nine dimers (**Figure 4.2a,b**), and the number of k -mers grows approximately exponentially with k (**Figure 4.2c**). While there are 27 permutations of trimer sequences, some cannot exist due to tree constraints, while some *LtS* trimers are included as they capture different structures than *StL* trimers. The *StL* trimers CTT and TTT do not occur as the latter T is a complete subtree that is smaller than its preceding sibling subtree. Any *LtS* k -mer with an A or T in the middle (of which there are 14 trimers), such as AAT or CCTC, describes a sequence of bifurcations not captured by any *StL* k -mer. In contrast, the AC *StL* dimer is equivalent to the TC *LtS* dimer as in both cases the C is the smaller-side child of an A, just as with the ACT *StL* and TCT *LtS* trimers the CT is the smaller-side child of an A.

The count and the proportion (i.e. count per sequence length) of each dimer are dependent upon node type proportions. For instance, a sequence with more C bifurcations than another would be expected to also have more CC dimers. The same effect applies to measuring trimer proportions with regard to dimer proportions, and so on. In order to account for these dependences, we normalize the analysis of k -mers for a given neurite sequence by a set of control trees with the same distribution of $(k-1)$ -mers. Specifically, a

set of 100 surrogate tree sequences is generated (sampled from the entire set of tree topologies) and constrained by ($k-1$)-mer proportions and tree size (**Figure 4.2d**). This produces an expected distribution of 100 values for each sequence and k -mer, allowing the computation of a percentile rank (PR) for the k -mer proportion in the original neurite sequence. The PR provides a value comparable across k -mers, arbor types, and individual sequences (i.e. axons, dendrites, and pyramidal apical dendrites). A k -mer that occurs significantly more often across a group of neurites than in the surrogate set is considered a motif; conversely, a k -mer that occurs significantly less often across a group of neurites than in the surrogate set is considered an anti-motif.

Percentile rank is calculated as

$$\frac{R - 0.5}{N}$$

where N is the number of values in the distribution, and R (rank) is the position of the value in an ordered list of values in the distribution. In case of multiple instances with the given value in the distribution, the rank is given by their central index (or the mean of the central indices) of instances with the value. For example, if the first 15 values are '3', then the rank of the value '3' will be 8 ($[1 + 15]/2$), and the PR will be 7.5%. If the value is smaller than all values in the distribution, the rank will be 1 and the PR will be $(1 - 0.5)/100 = 0.005$ or 0.5%. If the value is larger than all values in the distribution, the rank will be 100 and the PR will be $(100 - 0.5)/100 = 0.995$ or 99.5%.

All statistical tests used the Wilcoxon rank sum or signed rank tests. P-values determining whether a k -mer was significantly different than the baseline were adjusted for multiple testing using Bonferroni correction given the number of k -mers.

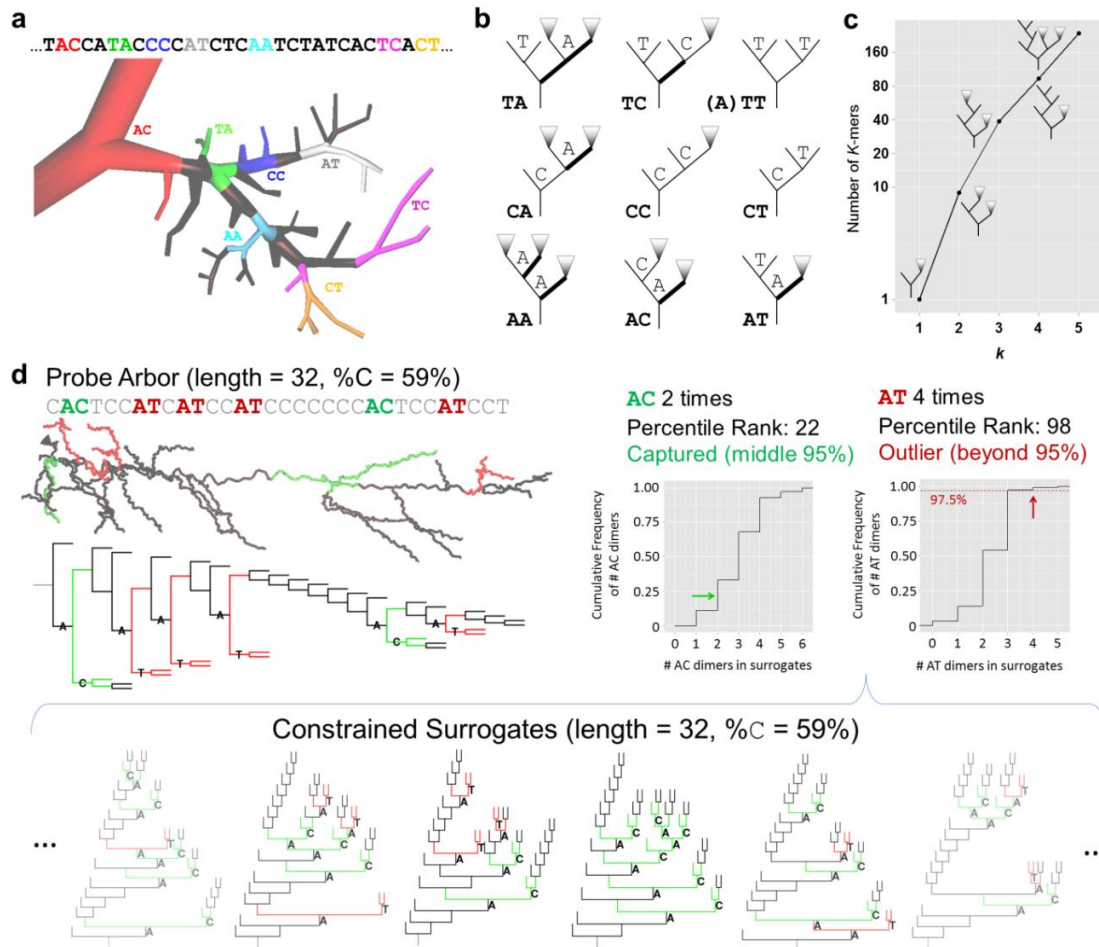


Figure 4.2. Measuring k -mers. **a.** Highlighted dimers in a portion of a fly tangential cell (Borst and Haag, 1996) and its associated sequence. **b.** Dimer schematics displaying the possible configurations. Triangles represent subtrees of unspecified size and shape. Bold segments indicate branches leading to the larger side subtree of the parent **A** node. Given the (Smaller then Larger) traversal method, the **TT** dimer must be preceded by an **A**. **TA** and **TC** schematics are examples, as additional bifurcations could be found between the parent **A** and small-side **T** nodes. **c.** Number of k -mers (with examples) by k shows an approximately exponential rise. **d.** Calculating the percentile rank of a k -mer given the distribution of k -mer counts in the source sequence's surrogate population. An example apical dendrite (NMO_02582 from [Nikolenko et al., 2007]), dendrogram, and sequence are shown along with cumulative distribution of k -mer counts for k -mers **AT** (red) and **AC** (green). **Below:** Six out of 100 node-type-constrained surrogates are shown. The example k -mers are highlighted and their counts compose the distributions. Colored dots show the respective percentile ranks of the apical dendrite k -mer counts, with **AT** being above nearly the entire surrogate distribution (thus constituting a motif) and **AC** being “captured” inside the middle 95% of its surrogate distribution.

Constrained tree-sequence generation

An algorithm for sampling length and node-type constrained surrogates was developed based on that used to generate size-constrained trees (Harding, 1971). However, no known method was found for rapid random tree generation fitting the constraints of k -mers with $k \geq 2$ while ensuring that the resulting trees are uniformly sampled over the distribution of tree shapes. The enormous number of trees that need to be generated before one matching the k -mer constraints is found made the length and node-type-constrained algorithm impractical. Therefore, multiple programs, one each for dimer, trimer, and tetramer constrained surrogates, were written that built tree shapes with k -mer defined components. The resulting distributions were compared with a smaller set of those produced by accepting constraint-satisfying trees from among those generated using the algorithm for producing uniformly distributed tree shapes (specifically the node-type constrained version). Minor modifications in the code were made to minimize the small deviations in $k+1$ -mer distributions seen in the “ k -mer built” surrogates (detailed comments provided within the code). The most extreme deviations were a median difference of 0.1% between constructed surrogate and constrained uniform surrogate proportions for trimers CTC and TTC.

Tree growth methods

Simple growth models were used to generate trees for comparison with available data. The models are versions of the general QS-model (Van Pelt and Verwer, 1986). The QS-model takes three parameters as constraints to stochastically generate a tree. A size parameter determines the number of branches. The parameter Q determines the propensity for bifurcation events to occur at terminal branches versus interstitial (non-

terminal) branches. A value of 0 produces purely terminal growth, a value of 1 produces purely interstitial growth, and values of 0.5 produces segmental (uniform) growth.

The parameter S determines the branch order bias of bifurcation probability. A value of 0 means no order bias; positive and negative values produce a bias towards low- and high-order branching, respectively. Specifically, the probability of a terminal node being selected to branch at a given branch event is given by $C \times 2^{-S\gamma}$, where C is a normalization constant and γ is branch order. The terminal growth model referenced in **Figure 4.4b,d** is low-order biased, using $Q=0$ and $S=0.4$. The segmental growth model used has no order bias, with $Q=0.5$ and $S=0$.

The NeuroMorpho.Org dataset

All data used in this study were downloaded from the NeuroMorpho.Org database (version 5.6), which houses neuronal reconstructions from a wide variety of species, brain regions, cell types, labs, and experimental methods (Ascoli *et al.*, 2007). In order to minimize confounds, we eliminated neurons cultured in non-organotypic environments ($N=29$) as well as the topologically simple neurons from the OpenWorm archive ($N=302$). Neurons from non-control (e.g. drug-affected) conditions were also eliminated ($N=1288$), and only one of multiple reconstructions found to be traced from the same neurons were kept ($N=25$ removed), bringing the dataset to 8,223 reconstructions.

Each reconstruction was partitioned into different neurite arbors (indicated by the ‘type’ column in SWC-formatted reconstruction files). Given the clear differentiation between axons and dendrites, the two arbor types were analyzed separately. Additionally, for pyramidal cells, apical dendrites were sequenced as a third distinct arbor type given

its well-known morphological differences from the basal trees (Whitford *et al.*, 2002; Donohue and Ascoli, 2008). Pyramidal basal dendrites were grouped with non-pyramidal dendrites, and pyramidal neurons without differential labeling of apical and basal arbors were removed from the dataset (N=20). Apical dendrites of non-pyramidal cells were ignored (N=11).

Many neuron types have multiple dendritic trees, sometimes with fairly few bifurcations each. Separate dendritic trees originating from the same soma were combined into a single binary dendritic arbor, substituting all initial segments (root stems) with connecting bifurcations. The assembly of individual trees into a joined arbor depended on the relative distances of their initial segments of each, with the nearest pair becoming siblings of a new bifurcation located mid-way between them, and the process repeated until only bifurcations remained. Although the same method can similarly handle multifurcations (i.e. 3 or more child branches), NeuroMorpho.Org preprocessing eliminates such cases by splitting them into multiple bifurcations at successive segments. Based on proximity of bifurcations, up to 1.5% of dendrites, 2.8% of axons, and 0.6% of apical dendrite branching events were trifurcations. A minority of datasets included reconstructions of spines labeled as short dendritic branches. To avoid confusing arbor topology with representation of spines, dendrites were further processed by removing any terminal dendritic branch shorter than 2 microns (969 dendritic arbors affected; branches removed: mean of 7.1, median of 2, maximum of 546). This preprocessing step had little impact on the conclusions of this study, but had a notable effect on analysis restricted to larger dendrites, as explained in the Results.

Given this work's focus on topological pattern detection and analysis, a substantial number of neurites were discarded due to having too few bifurcations and therefore insufficient complexity. Trees of a given size have a limited number of shapes (**Figure 4.3a**), that number increasing approximately exponentially with tree size (**Figure 4.3b**). We therefore picked a minimum tree size in order to maximize the available data while minimizing the chance of two unrelated trees having identical topology. At a size of 20 bifurcations, the odds of two trees in the dataset matching by chance (N=276) is about 5%; the odds halve at a size of 21 (N=280), with the chances still less than 25% if allowing for a single edit (insertion or deletion of a branch) from a tree of size 21. Using a minimum tree size of 20 bifurcations (inclusive), 1,056 axons, 2,460 dendrites, and 1,588 pyramidal apical dendrites were excluded. The final dataset contained 6,798 neurites in total, with 1,255 axons, 4,686 dendrites, and 857 pyramidal apical dendrites.

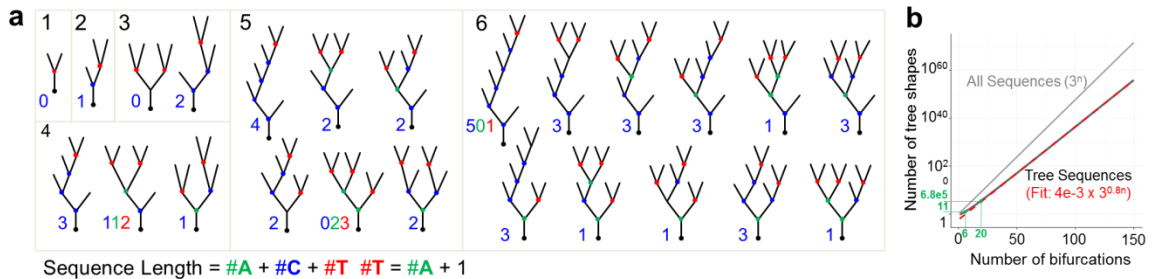


Figure 4.3. Tree size and complexity. **a.** Complexity of trees is limited by tree size. Here are shown the set of possible tree shapes for trees with 1 to 6 bifurcations. Additionally, the number of T nodes (red dots in sample trees) is always 1 more than A nodes (green dots). Thus, size and number or percent of C nodes (yellow dots) fully captures node-type statistics. **b.** Number of tree shapes for tree size (in # bifurcations, or sequence length). The relationship is approximately exponential, though the number is smaller than for the set of all possible sequences of the same size and alphabet unconstrained by “treeness”. Green lines indicate the 11 topologies of trees with 6 bifurcations displayed in (a) and the 680,000 tree shapes with 20 bifurcations which serves as the minimum complexity cutoff for the analysis.

The content of NeuroMorpho.Org is representative of the species, brain regions, and cell types that are commonly studied and reconstructed (Halavi *et al.*, 2012). After the above described selection, in the dataset used here Human data made up 37% of dendrites, Mouse 28%, and Rat 23%. Axons were made up of 50% Rat, 35% Mouse, and 11% *Drosophila*. Brain regions were represented primarily by Neocortex in both axons and dendrites, followed by Hippocampus and *Drosophila* Olfactory bulb in axons, and Hippocampus and Retina in dendrites. Apical dendrites came primarily from Rat (53%), Mouse (35%), and Monkey (10%), with 68% from the Neocortex and 31% from Hippocampus with nearly equal proportions between CA1 and CA3. Although the data is heterogeneous and further motif distinctions can be made between species and cell types of a given arbor type, this dataset reliably captures general research trends in terms of arbor type and metadata distribution, and thus serves effectively as a test bed for topological sequence analysis.

Results

Length and node-type proportions across arbors types

Many morphological features of neurons, particularly the arbor spatial coverage, total wiring length, and typical branch diameter, can be visualized and roughly estimated simply by looking at a reconstruction. Some topological features involving similarly straightforward calculations, such as the proportion of node types, are often quite difficult to discern visually (**Figure 4.4a**). Since the number of T nodes ($\#T$) is always one more than the number of A nodes ($\#A$), node type proportion information is reported compactly as the sequence length and proportion of C nodes (%C).

$$Length = \#A + \#C + \#T$$

Equation 4.1

$$\#T = \#A + 1$$

$$\#A = Length - \#C - (\#A + 1) = \frac{Length - \#C - 1}{2}$$

$$\%A = \frac{\#A}{Length} = \frac{1 - \%C - 1/Length}{2}$$

Equation 4.2

$$\%T = \frac{1 - \%C + 1/Length}{2}$$

Equation 4.3

Axons and dendrites show significantly different bifurcation type proportion distributions, measured by the proportion of C nodes (%C), with some overlap (**Figure 4.4b**), reflecting different functional properties and growth mechanisms. Pyramidal cell axon and dendrite distributions are broadly similar to the overall trends, but show a broader axonal distribution. Apical dendrites have %C values that fall between axons and (basal) dendrites.

The lower %C values of (non-apical) dendrites and the high values of the axons are suggestive of two conceptual growth models focusing on the site of bifurcation, namely terminal growth and segmental growth (van Pelt and Verwer, 1983). Terminal growth consists of a model neurite bifurcating only at terminal segments, reflecting growth cone bifurcation, for some predetermined tree size (see schematic in **Figure 4.4c**). A segmental growth model allows bifurcations to occur with equal probability at terminal and interstitial branches. The %C distribution of dendrites matches fairly closely to the

values of trees generated by a terminal growth model. Specifically, a small bias towards bifurcating at low-order branches (see **Methods**) is sufficient to achieve the precise overlap of low %C (**Figure 4.4b,d**). Conversely, axons more closely fit the segmental growth model.

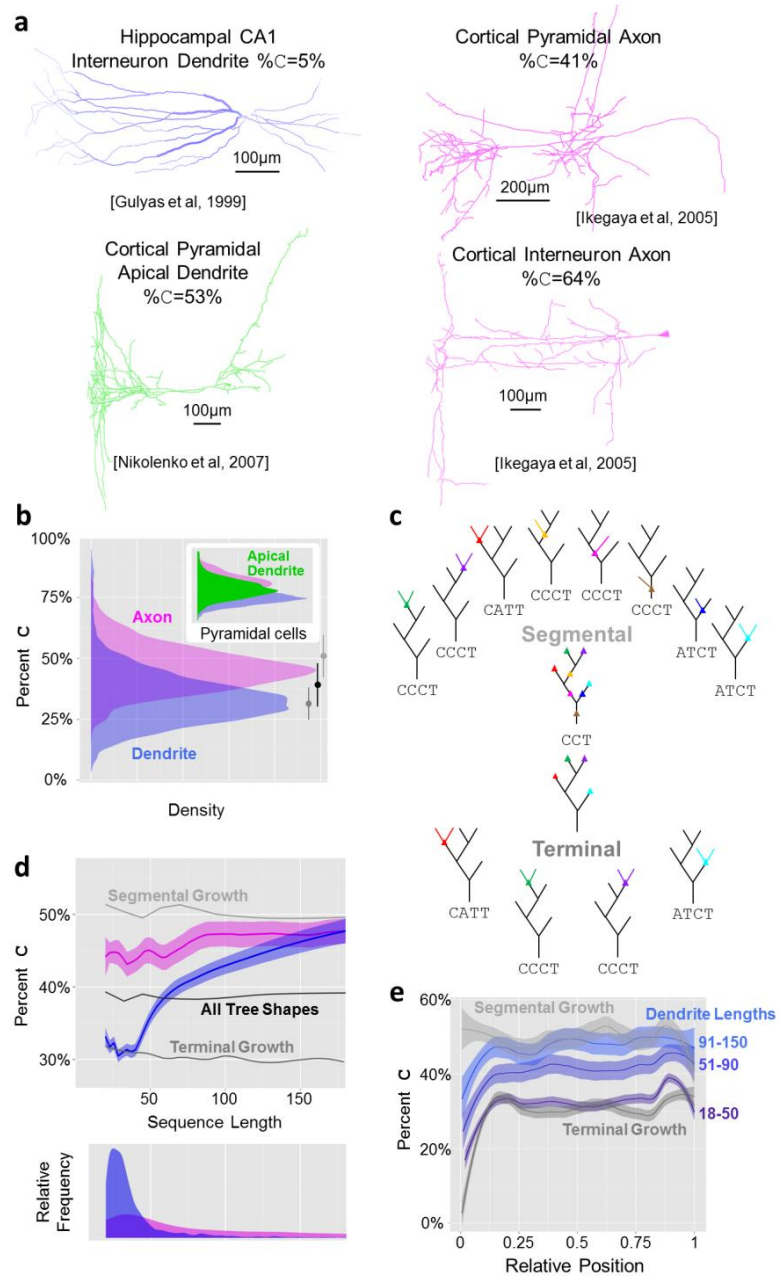


Figure 4.4. Neurite size and node type features. **a.** Example morphologies along with node-type proportions (%C) illustrate how difficult it is to estimate topological patterns by visual inspection of full morphologies (Gulyás *et al.*, 1999; Ikegaya *et al.*, 2005; Nikolenko *et al.*, 2007). **b.** Percent C distributions of axons (magenta) and dendrites (blue) overlap but are clearly distinct. **Inset:** Pyramidal cell basal apical dendrites (green) fall between basal dendrites and axons, which respectively are similarly distributed to the (non-apical) dendrites and axons of all (non-pyramidal) neurons. Biased terminal (dark gray) and segmental (light gray) growth bound the neurite populations, with the unbiased distribution of tree shapes (black) falling in between. **c.** Percent C versus sequence length for axons and dendrites compared to the tree shapes baseline, segmentally grown trees, and terminally grown trees with low-order bias. **Below:** Distribution of sequence lengths for axons and dendrites. Axons fall in between segmental and terminal growth while the

bulk of dendrites display terminal growth followed by a possibly segmental growth-based rise in %C with larger sequence lengths. **d.** Schematics of terminal (top) and interstitial (bottom) growth starting from a representative seed tree shape with sequence CCT. Colored dots represent potential bifurcation points given the growth mechanism, with their respective resulting branches seen in the trees surrounding the initial tree. **e.** Percent C as a function of relative position within sequences for dendrites of several sequence size groups along with segmental and terminal growth. The increase in %C with sequence length is sequence-wide and not specific to distal portions of trees. The initially low %C and rise to stability is similar to that displayed by terminally grown trees.

Most dendritic arbors have relatively few branches (**Figure 4.4d - bottom**), but for those that are larger than 40 bifurcations, the %C trends upwards until matching the %C distribution of axons at approximately 150 bifurcations (**Figure 4.4d**). This trend was systematic across all cell types consisting of a sufficient number of large dendrites (without removing likely spines from reconstructions, the trend had an even steeper slope and rose well beyond %C of axons). We initially speculated that this effect could be due to an increase in C bifurcations farther out in dendrites; however this hypothesis is proven false by analyzing the %C as a function of sequence position (**Figure 4.4e**). The only position effect, also observed in trees generated by terminal growth, is seen at the beginning and does not change substantially in larger dendrites. This suggests that larger dendrites are primarily larger by virtue of a segmental growth process and a greater proportion of C bifurcations.

Axons show only a very small change in %C with sequence length. This change could be an artifact of incomplete reconstructions. Axons generally cover far greater areas relative to dendrites and thus are commonly only partially reconstructed. The relative stability of the %C suggests that branching patterns might be largely unaffected

by partial reconstruction; however, as discussed below, dimer and trimer trends in axons do show a tree size effect.

Motifs and Anti-motifs

When analyzing k -mers of increasing length, it is expected that at some k there will be little if any additional information beyond the patterns already identified with shorter sub-sequences. In order to determine at what value of k the analysis should conclude, we identified the proportion of k -mers that were predicted by the $(k-1)$ -mer-constrained baseline for each k . If the PR fell within the middle 95% ($2.5\% < \text{PR} < 97.5\%$) then it was considered “captured” by the baseline. Trimer proportions almost fully constrained the representation of neurite tree sequences such that 99.1% of tetramers were captured (**Figure 4.5a**). As the capture rate of an additional set of trimer-constrained surrogate tetramers (distinct from the set used for normalization) was 99.4%, the gap between real neurite tetramers and the additional surrogates was $0.3 \pm 0.03\%$ (SEM). The additional k -mer constrained surrogate set was necessary due to the discrete nature of the k -mer distributions. Note that rather than a surrogate set achieving 95% of k -mers captured as would be expected with continuous data, the capture rate increases with k because more baseline count distributions have either 0 variance or the top or bottom of the distribution is composed of only a single value. The results are similar across arbor types at $k=4$, with axons showing the lowest capture rate and largest gap between neurites and baseline of $98.1 \pm 0.6\%$ and $0.6 \pm 0.1\%$, respectively. Although the difference between neurite and baseline was statistically significant due to the large n of over 6,000 sequences, the negligible effect serves to signal a cutoff for analysis beyond trimers. In

other words, describing the neurite distributions of monomers (%C), dimers, and trimers is also sufficiently informative of the distribution of longer patterns (capturing e.g. all but 0.3% of tetramers).

Except for one (CC), each of the 9 dimers is either a motif or an anti-motif, primarily reflecting terminal growth (**Figure 4.5b**), in stark contrast to the node type results (monomer pattern) observed for axons. Percentile rank values differ by a significant but small amount between axons and dendrites, with pyramidal apical dendrites deviating from baseline in the same direction but by far smaller amounts for each k -mer (**Figure 4.7**). Because k -mers for a given k are not independent of one another, a more detailed analysis requires a grouped approach. A k -mer that diverges from baseline substantially more than its counterbalanced k -mers, such as CA compared to AA and TA, is particularly illuminating. In this case, A nodes tend to descend from other A nodes, either on the small (AA) or large side (TA), and are particularly unlikely to be children of C nodes. Thus, axons and (non-apical) dendrites have highly arborizing regions topologically close to the soma, while the CT motif indicates that C nodes tend to occur primarily near neurite subtree terminations. In comparison, apical dendrites have much less extreme values for all dimers, better reflecting a segmental growth mechanism and a more even distribution of C nodes throughout the tree.

The strong CT motif and strong TT and AT anti-motifs are at least in part due to the greater number of growth pathways that produce trees with CCT or ATCT compared to ATT subtrees (**Figure 4.5c**). Increasingly complex terminal growth examples show even larger proportions of CT dimers relative to AT, while segmental growth show a

balance of the two while still maintaining low proportions of TT dimers. The CT/TT effect is consistently observed in dendrites and axons from a variety of neuron types, and particularly Martinotti interneurons (**Figure 4.5d-e**).

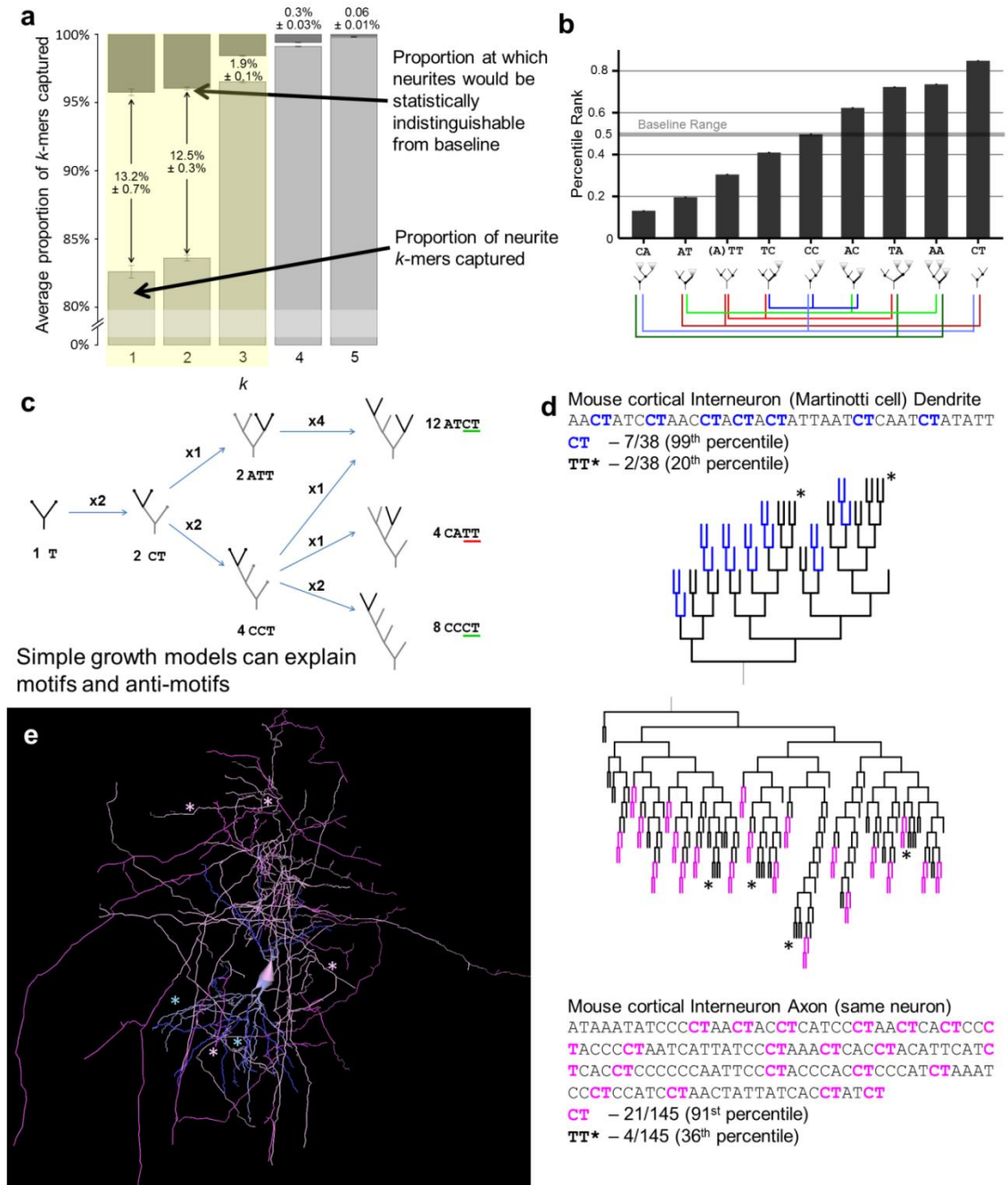


Figure 4.5. Dimer analysis reflects terminal growth effects. **a.** Average proportions of captured k -mers (percentile rank between 2.5th and 97.5th) at each k for all neurites, with a break between 0 and 80%. Over 95% of trimers are captured, as are over 98% of tetramers, suggesting that most analyses should focus up to dimers, and possibly trimers. Darker descending bars represent the baseline, or chance level based on bootstrapping of surrogates. Numbers between or above bars signify the gap in percent captured between real neurites and the baseline (i.e. statistical equivalence). **b.** All axon and dendrite dimers except **CC** are motifs or anti-motifs. Colored lines below dimer schematics show associations between dimers. When one dimer proportion increases (and in turn its percentile rank) another must decrease. **c.** Growth processes produce specific topological patterns. Starting with a tree of a single bifurcation, the tree grows one bifurcation at a time from terminal nodes. After two bifurcation steps the **CT** to **TT** ratio is 2:1 as there are twice as many ways a **CCT** shape can emerge compared to an **ATT** shape. After a third bifurcation (at right), the dimer ratio is 5:1, and still 3:1 controlling for %**C**. **d and e.** **CT** and **TT** motifs and anti-motifs shown in an exemplar interneuron's dendritic and axonal arbors (NMO_00340 from [Wang et al., 2002]) in sequence and dendrogram (**d**), and full morphology (**e**). Sequence and dendrogram highlighting indicate **CT** dimers in the axon (pink) and dendrite (blue). In the morphology, the darker color indicates the **CT** dimers. Asterisks (*) indicate the **TT** dimer in both representations. All error bars are standard error of the mean.

Trimer patterns generally continue several trends revealed by dimers including the nearer-to-baseline profile of apical dendrites and the similarity between dendrites and axons. Of the 39 trimers, dendrites and axons share 9 anti-motifs and 11 motifs. Trimer motifs CAT, CTA, CCT, and ATC, and anti-motifs AAT, ATA, CTC, TCT, CCA, all are consistent with primarily terminal growth and trees with C nodes associated with terminal T nodes. The AT dimer, a single branching event from a C node, is also seen as associating with C nodes in motifs CAT and ATC rather than with arborizing A nodes as evidenced by anti-motifs AAT and ATA.

A particularly distinguishing feature that sets axons and dendrites apart manifests in a small collection of trimers (TCT, ACC, ACT, TCA, and TCC) that show a tendency in axons for larger-side subtrees to continue rather than terminate (**Figure 4.6**). Specifically, while both axons and dendrites have CCT as a motif, only axons have TCT as strong anti-motif and ACT as a strong motif. TCT represents complete subtrees ATCT or ACTCT

(since the latter subtree must be of equal or larger size) and is counterbalanced by TCA and TCC which allow for larger-side subtrees of increased size. The ACC anti-motif is most simply explained as a counterweight for the ACT and TCC motifs. This difference between axons and dendrites is at least in part due to their different tree sizes, as these trimers are among the few that become more extreme with sequence length (consistent with terminal growth effects). However, even accounting for number of branches, some difference remains between axons and dendrites in these trimers, possibly related to the same functional or developmental mechanisms responsible for the different node type proportions.

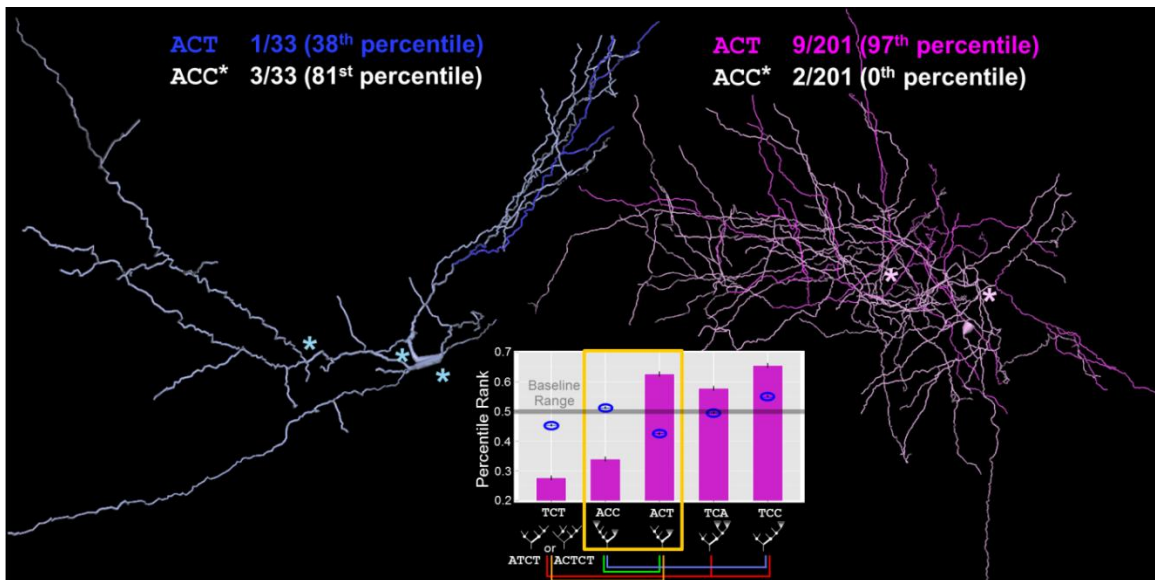


Figure 4.6. Distinguishing trimers of axons and dendrites. Axons and dendrites share some motifs consistent with dimers and their interpretations, but they differ on a small related set of k -mers. A rat hippocampal CA3 interneuron dendrite (left; NMO_00837 from [Chitwood et al., 1999]) displays a relative abundance of **ACC** trimers (asterisks) but only one **ACT** occurrence (dark blue). A rat cortical basket cell axon (right; NMO_07461 from [Kawaguchi et al., 2006]) has 9 occurrences of **ACT** (magenta) but only 2 of **ACC** (asterisks). **Graph Inset:** The **ACT** k -mer is a motif for axons (magenta bars) and a slight anti-motif for dendrites (blue ovals), while **ACC** is an anti-motif for axons and neither motif nor anti-motif for dendrites.

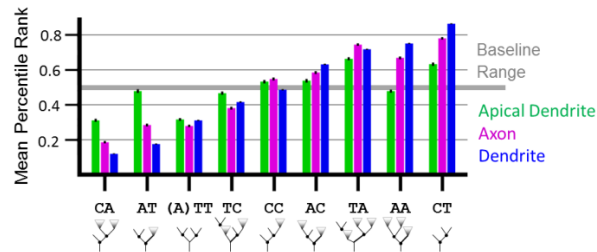


Figure 4.7. Dimer motifs and anti-motifs by arbor type. All mean dimer percentile ranks (with SEM error bars) are shown for each arbor type. Apical dendrites are closer to the baseline (50%) than axons and dendrites except for the case of (A)TT where all arbor types are approximately the same.

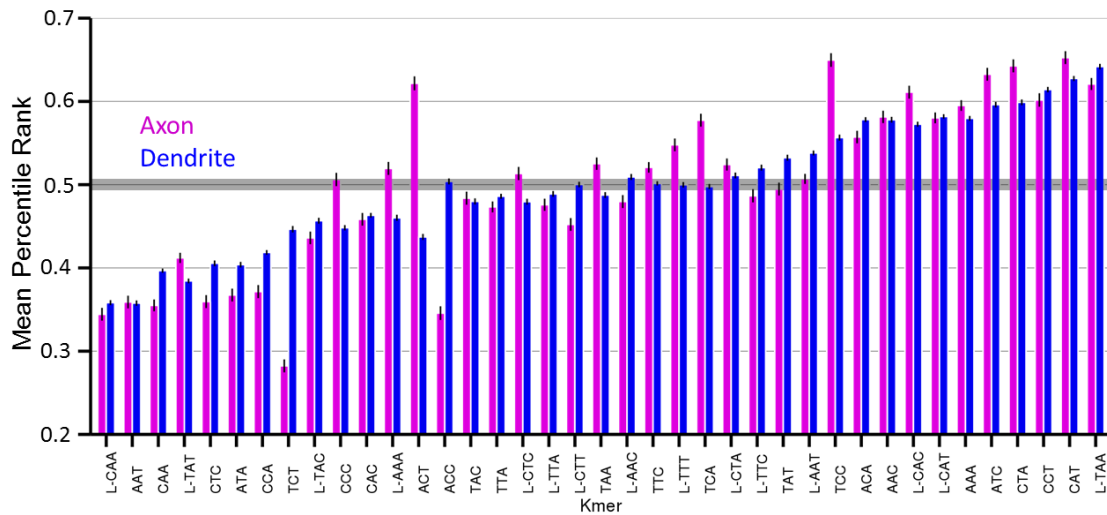


Figure 4.8. Axon and dendrite trimer motifs and anti-motifs. Axons and dendrites exhibit highly similar trimer patterns, with a few notable exceptions. Approximately one-third of trimers are motifs and another one-third anti-motifs.

Discussion

The novel method of representing neuronal trees as sequences of bifurcations introduced in this work facilitates detection and analysis of branching patterns. Using an exclusively topological approach we show that sequence length and subsequences up to three characters long almost completely capture the local topology of neuronal axons, dendrites, and apical dendrites.

Node type distributions highlight the most substantial differences between axons and dendrites, with axons having higher proportions of terminal side-branches (C nodes). The vast majority of dendrites exhibit particularly low proportions of C nodes, but the proportion increases considerably for dendrites with greater than 40 bifurcations. Dimer and trimer motifs show much greater similarity than difference between axons and dendrites, highlighting a primarily terminal growth effect with some particular and likely complex segmental growth process occurring in axons and larger dendrites. Pyramidal neuron apical dendrites (mainly cortical), on the other hand, stand out as having a k -mer profile very similar to a basic segmental growth model with relatively small deviation from the baseline of tree shapes.

Dendrites have generally denser spanning fields than axons and the functional role in processing inputs from other neurons can further constrain their shape. While homogenous targeting can lead to winding and asymmetric trees in models that only minimize neurite wiring, efficient signal integration (i.e. current transfer to the soma) demands shorter paths, which in turn require more symmetric trees (Cuntz *et al.*, 2007). It is tempting to hypothesize that terminal growth processes are simpler or more efficient at producing dendrites satisfying such constraints, and thus the dendritic motif profile

reflects those dominant features. Pyramidal apical dendrites, particularly those in the neocortex, have a more heterogeneous spanning field, with relatively sparse coverage along the trunk and denser coverage in the distal tuft (Spratling, 2002; Spruston, 2008). Segmental growth may better fit such a target region shape, though biased retraction following initial terminal growth, which is known to take place for some differentially projecting pyramidal subtypes (Koester and O’Leary, 1992), is an additional factor that impacts topology (van Pelt, 1997). The molecular and biophysical specialization of the apical dendrite (Whitford *et al.*, 2002; Matsuki *et al.*, 2010) optimizes input integration by minimizing the effective path distance of farther synapses (Vetter *et al.*, 2001).

Axon motifs indicate a more complex growth process with dimers and trimers suggesting primarily terminal growth but node-type proportions reflective of segmental growth. While axons are not as constrained as dendrites in terms of path distance and target region size, they are the neuronal component most responsible for navigating through the environment to create connections. Axons are also characterized by a more substantial energy footprint and specific energy regulation mechanisms (Yang *et al.*, 2015). Developmental studies have shown that axons branch interstitially (Kalil and Dent, 2014); however, during the interstitial branch outgrowth the growth cones stop (Bastmeyer and O’Leary, 1996; Szebenyi *et al.*, 1998). In conjunction with axonal self-avoidance (Sagasti *et al.*, 2005; Grueber and Sagasti, 2010) and retraction during development (Luo and O’Leary, 2005), this complex orchestration might produce terminal growth-like features.

Axon morphology is also impacted by post-developmental activity and plasticity. Axons are motile and plastic, displaying branch growth and retraction in adult animals (De Paola *et al.*, 2006; Stettler *et al.*, 2006). While this is also true of dendrites, on average axons have much greater inter-branch distance along which to bifurcate or move in response to environmental cues. These capabilities have important implications for network plasticity and learning, including the possibility of branch outgrowth in response to local network activity. Such branching would be segmental in nature since it could occur at any location along an axon; such a mechanism would at least partially explain the high %C in axons, which otherwise display topological patterns indicative of terminal growth.

The motif results of axons will benefit from confirmation with a larger set of complete reconstructions, as many axons in the dataset are likely incomplete (Parekh *et al.*, 2015). However, the relationship between k-mer percentile ranks and sequence length generally followed the trends seen in a basic terminal growth model. Assuming axon tree size as a proxy for completeness, it is likely that the general findings will be confirmed. Nonetheless, subsets of the axon data may reflect reconstruction artifacts.

Data biases

Given that our analyses primarily focus on arbor types across cell types and brain regions, the results are susceptible to bias in the dataset. Our exploration of potential distinctions between cell types in dendrites revealed few clear differences beyond sequence length distributions given the sample sizes available. The same appears to hold for axons, though the sample size of diversity of neuron types is far more limited.

Analysis in the companion study do show substantial global differences between four axon subtypes, with only modest motif profile differences between three of them (Gillette *et al.*, 2015). This suggests that the results are generally applicable for dendrites and axons, and that most differences between classes, while potentially interesting and worth further study, are small compared to deviations of the arbor types from the baseline.

Apical dendrites, on the other hand, show substantial differences between brain regions. The results presented here are dominated by the more numerous neocortical pyramidal cells. The hippocampal CA1 and CA3 motif profiles differ from the neocortical profile and from each other, consistent with known morphological and anatomical differences between the 3 regions.

Alternative encodings and traversals

Although the current work focused on purely topological encoding, the approach to represent arbors as strings of characters can be extended with additional geometric characteristics of branches. Several branch-level features might be discretized and encoded, such as branch length, tortuosity, or bifurcation angle. With an expanded sequence based on a larger alphabet of characters, the number of k -mers would increase dramatically. Motif analysis could allow us to determine whether certain branch-level features co-occur with each other and with specific topological patterns, providing novel observations for forming new growth and functional hypotheses.

Classification is another potential use of any new representation and measure. In the case of topological motifs the variability is too large to be particularly useful in classification; however, other analyses using this representation as well as an expanded

encoding have the potential to aid in classification. A classification example and candidate encodings are discussed further in the companion paper (Gillette *et al.*, 2015).

Rather than a depth-first traversal, a tree could alternatively be encoded as a collection of sequences, each representing a path from root to tip (Basu *et al.*, 2011). The utility of the representation likely depends upon the specific investigative aim. One potential draw-back is that low-order branches will be represented multiple times, which, in addition to imposing a bias for certain types of analysis, multiplies the size of the representation. A benefit of such a representation lies in preserving the sequential pattern of all paths.

Model validation and additional applications

The current topological representation, along with expanded representations, could prove valuable as a more sensitive measure of emergent topology relative to standard morphometrics for gauging the quality of computational simulations of developmental processes. Standard topological metrics are useful for determining whether a particular model matches the true distribution, but motifs may provide a clearer indication of how the model and true neurite distributions diverge and what modifications might be necessary to improve the model. Experiments focusing on functional constraints, such as wiring efficiency or distribution of synaptic targets, would have the added benefit of more clearly associating motifs with neuronal function. Though not detailed here, the distinctive motifs of neurites from different neuron types suggest that topological arbor patterns may be impacted by the specific functions and anatomic contexts of neuron types.

Conclusions

The proposed sequence representation of binary trees is effective for quantifying the topological patterns of neuronal arbors. Motif analysis offers a measure to validate or refine models and hypotheses. Motifs analysis reveals that the local topological features of both axons and non-apical dendrites are similarly consistent with terminal growth-like processes despite substantial different overall size and function. In contrast, pyramidal apical dendrites exhibit a motif profile indicative of segmental growth.

The methods described here and in the following paper can generally be applied to any tree structure of sufficient complexity. Research in glia (Matyash and Kettenmann, 2010; Hartenstein, 2011), brain vasculature (Wright *et al.*, 2013), botany (Segura *et al.*, 2008), rivers/watersheds, and phylogenetics are all potential targets of a sequence analysis approach. Moreover, while the topological sequence representation is powerful as it is, it also can serve as a basis of other morphological features.

Acknowledgements

We thank Christopher L. Rees, Charise White, Alexander O. Komendantov, and Ruchi Parekh for their thoughtful feedback on an earlier version of the manuscript. This work was supported by NIH grant R01 NS39600.

CHAPTER FIVE: TOPOLOGICAL CHARACTERIZATION OF NEURONAL ARBOR MORPHOLOGY VIA SEQUENCE REPRESENTATION. II. GLOBAL ALIGNMENT

[An edited version of this chapter is in press: Todd A. Gillette, Parsa Hosseini, Giorgio A. Ascoli: in BMC Bioinformatics (2015)]

Abstract

Background: The increasing abundance of neuromorphological data provides both the opportunity and the challenge to compare massive numbers of neurons from a wide diversity of sources efficiently and effectively. We implemented a modified global alignment algorithm representing axonal and dendritic bifurcations as strings of characters. Sequence alignment quantifies neuronal similarity by identifying branch-level correspondences between trees.

Results: The space generated from pairwise similarities is capable of classifying neuronal arbor types as well as, or better than, traditional topological metrics. Unsupervised cluster analysis produces groups that significantly correspond with known cell classes for axons, dendrites, and pyramidal apical dendrites. Furthermore, the distinguishing consensus topology generated by multiple sequence alignment of a group of neurons reveals their shared branching blueprint. Interestingly, the axons of dendritic-targeting interneurons in the rodent cortex associates with pyramidal axons but apart from the (more topologically symmetric) axons of perisomatic-targeting interneurons.

Conclusions: Global pairwise and multiple sequence alignment of neurite topologies enables detailed comparison of neurites and identification of conserved topological features in alignment-defined clusters. The methods presented also provide a framework for incorporation of additional branch-level morphological features. Moreover, comparison of multiple alignment with motif analysis shows that the two techniques provide complementary information respectively revealing global and local features.

Introduction

Neuronal morphology has been an important research topic due to its relevance to neuron growth, electrophysiology, classification, connectivity, and pathology. In recent years the availability of morphological data has dramatically increased across a broad set of species, cell types, and conditions thanks in part to search and curation efforts of NeuroMorpho.Org (Ascoli *et al.*, 2007). Meanwhile, new staining, imaging, and reconstruction technologies are producing orders of magnitude more data from animal models in projects such as FlyCircuit (Chiang *et al.*, 2011) and FlyLight (Jenett *et al.*, 2012). Ongoing advancements are continuously improving automated reconstruction (Donohue and Ascoli, 2011; Parekh and Ascoli, 2013) and groundbreaking large scale endeavors such as the BRAIN Initiative are expected to expand the amount of data from mammalian organisms by additional fold factors. Importantly, these large datasets provide opportunity to validate universal morphological characteristics (Teeter and Stevens, 2011; Cuntz *et al.*, 2012) as well as to discover or test features and relationships of specific neuron types (Polavaram *et al.*, 2014). Drawing inspiration from the field of

genomics, where the Human Genome Project ushered in an analogous flood of data, a variety of new methods are being developed to enable search and analysis of the growing pool of morphological data.

Historically, most metrics devised for morphological analysis have focused at the whole neuron level or averaged across branches, with some derived methods quantifying a feature as a function of distance from the soma (i.e. Sholl analysis [Sholl, 1953; Scorcioni et al., 2008]). Pairwise comparisons have greater potential to detect fine structural distinctions, making them appropriate for search, curation, classification, and exploration of prototypical features. Recently several studies have used pairwise comparisons in morphological analysis. Pairwise sequence alignment of neurite path directional vectors helped determine the unique neuroblast lineage of larval *Drosophila* neurons (Cardona *et al.*, 2010). The NBLAST algorithm leveraged a form of vector comparison, with nearest edges in two neuron images aligned and measured by their tangent vectors and spatial distance (Costa *et al.*, 2014). Thus, anatomical position and overall shape were applied for search, clustering, and classification to a database of over 16,000 *Drosophila* neurons. The Path2Path algorithm compares neurons by assigning every path, from root to tip, of one neuron to the other. The distance is given by the deformation of the paths, modulated by the difference in topological hierarchy of points along the paths (Basu *et al.*, 2011). An extension of the Elastic Shape Analysis Framework captures the difference between trees based on path shape and topology, as well as bifurcation locations and angles (Mottini *et al.*, 2014). This method can also generate a representative “mean shape”, though the examples primarily represent

common path features. BlastNeuron, the most recent entry into the field, focuses on aligning branches both by topology and path shape via dynamic programming after first searching for similar neurons on the basis of morphometrics (Wan *et al.*, 2015). In addition to providing an efficient approach for search in large databases, the alignment component could prove useful in detecting and pinpointing differences between related neurons and between reconstructions of the same neuron produced by multiple algorithms, enabling error correction and even synthesis of those algorithms.

The tree edit distance (TED) compares the topology of two trees by determining the minimum sequence of edit operations required to transform one tree into another (Zhang, 1996). Specifically, each branch of two trees is aligned to a branch in the other tree or labeled as an insertion. Branch features such as length, volume, surface, and bifurcation angle can be represented; in this case an edit cost based on their differences is applied for each branch assignment. The TED has been used on tree structures in multiple fields (Gillette and Grefenstette, 2009; Heumann and Wittum, 2009) and constitutes the most related algorithm to what we present here.

We present an original strategy to evaluate alignment of topology distinctly from other branch features across a broad range of neuronal classes. Our method exploits the novel encoding of neuron trees as sequences of characters representing bifurcations presented in the preceding companion paper (Gillette and Ascoli, 2015). We align the resulting strings with a custom-developed Python package introduced here: Pattern Analysis via Sequence-based Tree Alignment (PASTA). We used model-based cluster analysis on alignment scores to group similar neurites. Furthermore, we generated a

consensus representation of clustered neurites by multiple sequence alignment revealing the conserved structural features of the corresponding trees. Sufficiently large neuron classes, well-defined by available metadata, were compared to the clusters to determine whether those classes are topologically distinct and, if so, what their defining global features are. Each arbor type of axons, dendrites, and pyramidal cell apical dendrites showed clear topology alignment clusters with distinctly conserved features. Moreover, we show that multiple alignment consensuses and motif analysis provide complementary levels of analysis of neurite topology. As an immediate application, the PASTA tool also enabled detection of previously unidentified duplicate reconstructions in the NeuroMorpho.Org database; this important curation step will be incorporated in the regular data processing pipeline of this repository. At the same time, the approach is extensible to more complex representations as required by the research goal.

Methods

Sequences are generated from neuronal trees as presented in the companion paper (Gillette and Ascoli, 2015): each branch is encoded as an A, C, or T depending on whether its child branches both bifurcate, bifurcate and terminate, or both terminate. A depth-first traversal keeps child branches relatively near to parent branches in the sequence representation, and traversing the smaller subtree (in terms of number of branches) before the larger (i.e. *StL* traversal) further preserves that proximity. This choice also allows for a more intuitive description of changes to a tree structure (**Figure 5.1a**).

Custom sequence alignment algorithms were developed for tree-derived sequences given their particular structural features. A global alignment algorithm was developed with specialized character types and alignment rules associated with bifurcation types. The scores produced by this method were normalized by size and a random-tree baseline prior to analysis. A multiple sequence alignment algorithm was also custom-adapted for tree-sequences from the genomics global and multiple alignment techniques.

morphologies of hippocampal CA1 and CA3 pyramidal cells (NMO_00588, 07558, 00219, 00213 from [Ishizuka et al., 1995; Megías et al., 2001; Michaelsen et al., 2010]) and their topology dendrograms illustrate alignments between neurites of the same sub-region along with normalized distances for all 6 pairs. **Inset:** Within- and between-group distributions of normalized distances (with mean \pm standard deviation) indicate that the pyramidal apical dendrites are topologically more similar to apical dendrites of their own group than to those in the other group.

Other well-established data manipulation and analysis methods utilized in this work are briefly described as well, such as the creation of a metric space from alignments for further analysis including classification and model-based clustering. Clusters are then statistically analyzed for associations with known cell groups. This Methods section also describes the (publicly available) dataset used in this study and related selection criteria. All code developed for this study (and associated documentation), including tree and sequence generation (Java), sequence alignment (Python), and analysis (R), is available open source at <http://krasnow1.gmu.edu/cn3/NeuriteSequence/>.

Global tree-sequence alignment

Powerful bioinformatics approaches exist for comparing gene sequences to each other to determine similarity and derive evolutionary relationships. A gene may mutate by conversion of one base pair to another (edit), the introduction of a new base (insertion), or the elimination of a base (deletion). Analogous operations allow the comparison of neuronal trees: the modification of branch properties, outgrowth of new branches, and branch retraction. While neuronal trees do not explicitly share an ancestry or phylogenetic relationship typical of nucleotide sequences, they do share growth programs based on neuron type and associated genetic expression. From a functional perspective, neuronal branches have properties that impact charge transfer, signal

integration, and space occupancy. The assembly plan of these features may indicate growth rules and functional attributes of the neuronal arbor as a whole.

Given the described encoding scheme (Gillette and Ascoli, 2015), the growth of a single neurite branch results in one of two possible sequence edits, with branch retraction expectedly producing the opposite edit in either case. Growth prior to a bifurcation, or on the child branches of a T bifurcation, yields the insertion of a C (transition between first and second trees/sequences in **Figure 5.1a**). Growth on the terminal branch of a C bifurcation, however, will turn the C into an A bifurcation and insert a T (transition from the third to the second tree/sequence in **Figure 5.1a**). Conversely, the retraction of a T's terminal branch deletes the T and turns the parent A into a C. If the T's parent is a C, then the retraction effectively deletes the C (the C bifurcation becomes a T and the original T disappears). A sequence traversed by the smaller subtree first ensures that the T insertion/deletion occurs adjacent to the A/C conversion.

The global alignment of tree sequences is modified from the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to account for the branching structure of binary trees (see **Appendix One**: for complete details), with alignment rules associated with bifurcation types (**Figure A1.1**). Alignment scores are given by the number of matching positions minus the number of gaps (**Figure 5.1b**). An additional penalty of -3 is applied for each region of consecutive gaps. This cost helps distinguish more closely related sequences and reduces the otherwise high probability of a perfect score for a pair of sequences of very different lengths. Perfect scores are problematic as they produce distance values of 0, making alignment-space embedding impossible. Alternative gap

opening costs were tried with no meaningful impact on the results reported here; however, tuning this parameter might help optimize future analyses depending on the specific questions of interest. Non-unitary gap costs have no impact (given the fixed scale of gap open costs) due to all matches adding the same amount to the score.

In order to focus on topological structure differences and not tree size differences, we normalized scores by sequence length. Specifically, a per-character score is computed and normalized against a length-pair matched random tree alignment baseline prior to analysis (**Figure A1.2**). The normalized alignment score of a pair of trees is then converted to distance by subtracting it from 1, the maximum possible score (see **Appendix One**: for complete details).

Spatial embedding and classification

In order to investigate the pairwise relationships among neuronal tree sequences, neurites were embedded in a multidimensional space consistent with alignment distances using Multidimensional Scaling (MDS) (Härdle and Simar, 2007). While edit distances would allow for standard MDS, the triangle inequality does not hold for tree alignment scores (Bille, 2005) or our normalized distances, thus a non-metric version was employed (Venables and Ripley, 2002). Instead of producing a space that preserves the original distance values in the embedded space, non-metric MDS seeks to preserve the relative proximity of the data elements. In other words, neurites that are closer to each other than to a third item in the space have that same relationship in the original data.

The embedded data were subjected to cluster analysis to determine whether different neuron classes can be distinguished in terms of their topological sequences. A

wide variety of clustering methods exist, including agglomerative clustering, k-means clustering, and model-based clustering. Agglomerative clustering, including Ward's method (Ward, 1963), can be run directly with distance values, obviating the need for MDS, but top level clusters are less stable than in other methods and visualization is limited. K-means assumes that the clusters are spherical and approximately of the same size (Steinhaus, 1956), but neither of those assumptions is sound with regards to the diverse population of neurite shapes. Model-based clustering was chosen due to its ability to test a variety of potentially optimal model types based upon the structure of the dataset (Fraley and Raftery, 2002; Fraley *et al.*, 2012). We used this approach to fit multivariate Gaussians with parameters for size, shape, and correlation between dimensions while varying the number of clusters. The simplest model with the greatest likelihood was selected using the Bayesian Information Criterion. Associations between clusters and known neuron classes were quantified by the relative representation of a given metadata group within a cluster and vice versa using the Bonferroni-corrected χ^2 test (see **Appendix One:**).

To determine whether the alignment space captured the same or additional information relative to that of standard topological metrics, we compared the corresponding abilities to correctly identify known neuron types using unsupervised and supervised classification. Model-based clustering is unsupervised and was restricted to two clusters for the purposes of classification. Linear discriminant analysis (LDA) is a supervised approach that produces a trained classifier based on a subset of the data (Dempster *et al.*, 1977; Weihs *et al.*, 2005). By applying each method with different

numbers of variables, an optimal classifier can be produced with the most informative variables identified. Classification accuracy was computed as the proportion of neurites in the appropriate class (additional details in Supplementary Material 2.3). Clustering performance was further measured by the adjusted Rand Index (aRI) (Rand, 1971; Hubert and Arabie, 1985). An aRI score of 1 indicates perfect cluster partitions along known classification, while a score of 0 indicates random partitioning of the neurites. The metrics tested were number of bifurcations, maximum branch order, average partition asymmetry, and caulescence (main path prominence) (Brown *et al.*, 2008).

Multiple sequence alignment

The multiple sequence alignment (MSA) process involves several steps: the creation of a composite allowing for registration of the multiple alignment, multiple alignment of all sequences to the composite, generation of a position-specific scoring matrix (PSSM) based on the prior MSA, iteration until the MSA is stable, and finally thresholding to extract a consensus sequence (**Figure A1.3a**). The composite is initially formed by aligning one sequence to another and then creating a sequence out of the alignment. Positions in which a gap exists are filled with the character on the non-gapped side of the alignment. This process is repeated with the growing composite now being aligned to each sequence in the set one at a time (**Figure A1.3b**).

The multiple alignment is formed by aligning every sequence to the complete composite in the same manner as was used in the pairwise global sequence alignment. This process identifies the sequence co-alignment positions without requiring an explicit score, thus no penalty is needed for initiating a gap region. Since the initial multiple

alignment could be suboptimal, a PSSM is generated with the positions defined by the composite and the scores defined by the number of sequences aligned at each position (**Figure A1.3c**). The multiple alignment is run again using the PSSM, the process iterating until stability is reached or the thresholded consensus fails to increase in size or average conservation.

Final consensus can be determined for any given minimum conservation threshold. Higher thresholds produce shorter consensus with higher average conservation and vice versa. A threshold of 50% was used for all MSA in this paper.

NeuroMorpho.Org dataset

The dataset used in this research is the same as in the companion motif analysis (Gillette and Ascoli, 2015), drawing from NeuroMorpho.Org (Ascoli *et al.*, 2007) version 5.6. Briefly, it was composed of all unique neurons of the control, non-cultured condition, except for pyramidal neurons without distinct apical/basal trees. Axons, dendrites (including pyramidal basal dendrites), and pyramidal apical dendrites were treated separately, with dendrites having any terminal branches smaller than 2 microns removed to avoid confusing likely spines for dendrites. Neurites with fewer than 20 bifurcations were eliminated to ensure sufficient complexity and reduce the probability of unrelated neurites achieving high similarity by chance. A total of 6,798 neurites were analyzed, including 1,255 axons, 4,686 dendrites, and 857 pyramidal apical dendrites.

Results

Using the separated arbors of neurons from the NeuroMorpho.Org database, each neurite was encoded as a sequence (Gillette and Ascoli, 2015), aligned with every other

sequence in the dataset, and the scores normalized and converted into distances (**Figure 5.1b**). Inspection of normalized distances from hippocampal CA1 and CA3 apical dendrites, two groups known to have distinct topologies, showed greater distances between sequences from different groups than from within the same group (**Figure 5.1c**). This verifies that the tree-sensitive alignment of neurite sequences can distinguish different neurite classes. We next undertook an unsupervised cluster analysis to identify groups of neurites for inspection of topological features.

Topology alignment space

An abstract sequence alignment space of neural trees was generated for the purposes of analyzing topological relationships between neurite types. The calculated distances were used to embed neurites into a feature space, hereon referred to as alignment space, using non-metric multidimensional scaling (MDS). Eight dimensions were sufficient for the entire dataset, with the first four showing clear structure in terms of density distribution (two-dimensional projection shown in **Figure 5.2a**). The spaces produced for each data subset are highly similar in the dominant first few dimensions which capture most of the original distance relationships. While the space is an abstraction, roughly speaking the top dimensions are mainly a function of the number of bifurcations and asymmetry. The first effect may initially seem counterintuitive since the number of bifurcations (sequence length) was explicitly controlled for; however, it is in fact consistent with the observation that dendritic branch type composition depends on tree size (see **Figure 4.4d,e**). At any rate, the correspondence between alignment dimensions and topological features is far from exact. Within a restricted region of the

alignment space, moving in a set direction will result in trees with increasing or decreasing richness in some structural elements; over broader spans of the abstract space, however, the directional gradient of different elements will typically vary.

The density distribution of the alignment space suggests groupings of neurites that are more related to each other than they are to other groups of neurites. Model-based clustering methods were applied to choose the best model and the optimal number of clusters, and then to detect the clusters. Clusters take the form of multivariate Gaussians with defined size, shape, and orientation, where parameters may be the same for all clusters or unique for each.

Analyzing all neurites in the dataset yielded seven clusters with all parameters specified and distinct. Of those, four were solely associated with dendrites, one with axons, one with pyramidal apical dendrites, and one principally with axons but also significantly with apical dendrites (**Figure 5.2b**). This is consistent with the greater number and diversity of dendrites in the NeuroMorpho.Org dataset. Accuracy (proportion of neurites falling into a cluster associated with the corresponding type) was 80% for dendrites, 65% for axons, and 54% for apical dendrites due to the significant overlap with one of the axon clusters.

Each arbor type was also analyzed separately to identify potential sub-clusters and related associations with known metadata. In the case of apical dendrites, the resulting three sub-clusters clearly associated with brain regions, specifically neocortex, hippocampal CA1, and hippocampal CA3 (**Figure 5.2c**). This result highlights the importance of clustering subsets, particularly when there is a bias in proportions of

available data. In some cases differences may exist but be too small relative to topological variability or have too small of a sample size to detect.

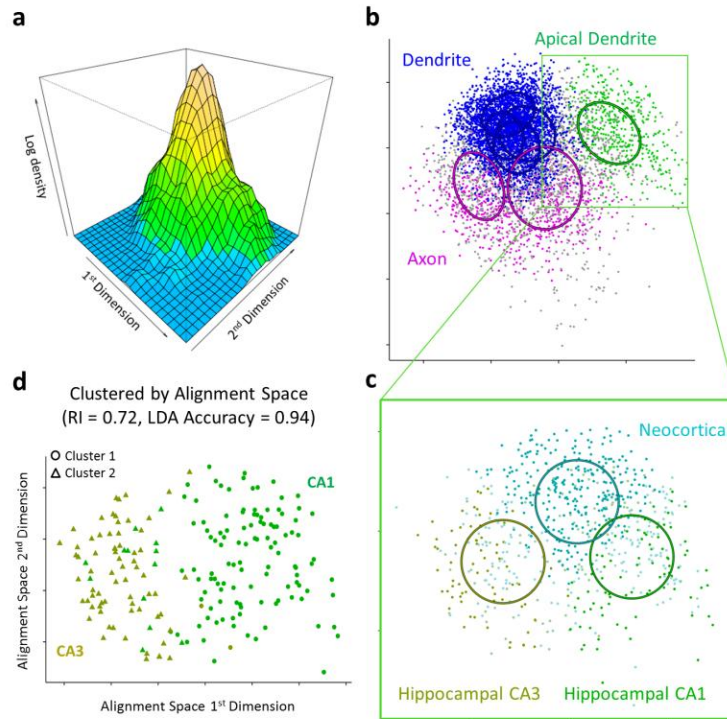


Figure 5.2. Alignment space and clustering. **a.** The log density plot of all neurite-derived sequences on the first two dimensions of alignment space suggests differentiable groups. **b.** Clustering of the alignment space reveals four clusters associated with dendrites (blue), two associated with axons (magenta), and one associated with apical dendrites (green); most neurites grouped with the cluster associated to their arbor types, but some classified with another cluster (gray). Ellipses reflect the covariance matrices of the cluster models. **c.** Sub-clustering of apical dendrites yields additional spherical clusters that associate with neocortex (cyan - top), hippocampal CA1 (green - right), and hippocampal CA3 (yellow - left). Apical dendrites improperly classified have lower color intensity. **d.** Unsupervised model-based clustering of the alignment space classifies CA1 and CA3 apical dendrites with an adjusted Rand Index of 0.72. A trained linear discriminant analysis achieves 94% accuracy.

Classification and comparison to morphometrics

In order to determine how well, relative to morphometrics, alignment space can distinguish neurites of different types, we used both model-based clustering and linear

discriminant analysis (LDA) to classify neurites. In the case of CA1 and CA3 hippocampal apical dendrites (**Figure 5.2d**), the optimal set of alignment space dimensions performed (slightly) better than the optimal principal components of topological metrics using both methods. The clustering approach produced an accuracy of 92% (aRI of 0.72) using alignment space compared to 91% (aRI of 0.67) using the principal components of the topological metrics. Alignment space performed even better using LDA, producing an accuracy of 94% compared to 89% with topological metrics.

In the case of *Drosophila* olfactory axons and mouse cortical pyramidal axons, alignment space (model-based: 87%; LDA: 87%) also produced better classification scores than topological morphometrics (model-based: 82%; LDA: 81%) using both methods. Of nine other neuronal class pairs that were at least minimally separated, the alignment space performed as well as (to within a 1% accuracy difference) or better than topological metrics in 7 out of 9 cases. These results show that, while topological metrics generally provide a robust characterization of neuronal classes, the additional detail captured by the sequence representation and alignment is sufficient to improve classification for some known classes. For the explicit purposes of classification, supplementary (non-topological) morphometrics are certainly called for; however, expanded sequence representation and pairwise alignment incorporating non-topological branch features have the potential to be more effective than those features summed or averaged across the neuron.

Dendrite and axon clusters

We next looked at how dendrites sub-clustered by topological alignment.

Dendrites, which make up the largest arbor type subset of the data (N=5,411), fell into 6 clusters with Gaussians of varying cluster size, shape, and orientation (**Figure 5.3a**). Four alignment dimensions were required to distinguish those clusters. Upon considering dendrite classes with sufficient specificity and size, seven classes showed highly significant association with one or more clusters (**Figure 5.3b**). The 6 dendrite clusters clearly separated cortical neurons from sensory neurons of several different species. The only class that reciprocally associated with only one cluster (and with a very high level of significance) consisted of primate neocortical (layer 2/3) pyramidal basal dendrites (cluster 1). Rodent neocortical pyramidal basal dendrites alone associated with cluster 2 and jointly with rodent neocortical interneuron dendrites in cluster 3. Clusters 4-6 associated with sensory and motor neurons. Motoneurons, primarily from cat, rat, and mouse, as well as rodent retinal ganglion cells, associated with cluster 4. The rodent retinal ganglion cells associated most significantly with cluster 5 along with adult fly tangential cells (visual system) and larval fly sensory cells. The latter two groups also associated with cluster 6. Thus, the three types of sensory neurons co-clustered, suggesting that topological features are shared by neurons which respond directly to the external environment. Tree size clearly distinguishes some of the clusters exemplified in some sample dendrites from each group, but differences in the number and distribution of major branches are also evident (**Figure 5.3c**).

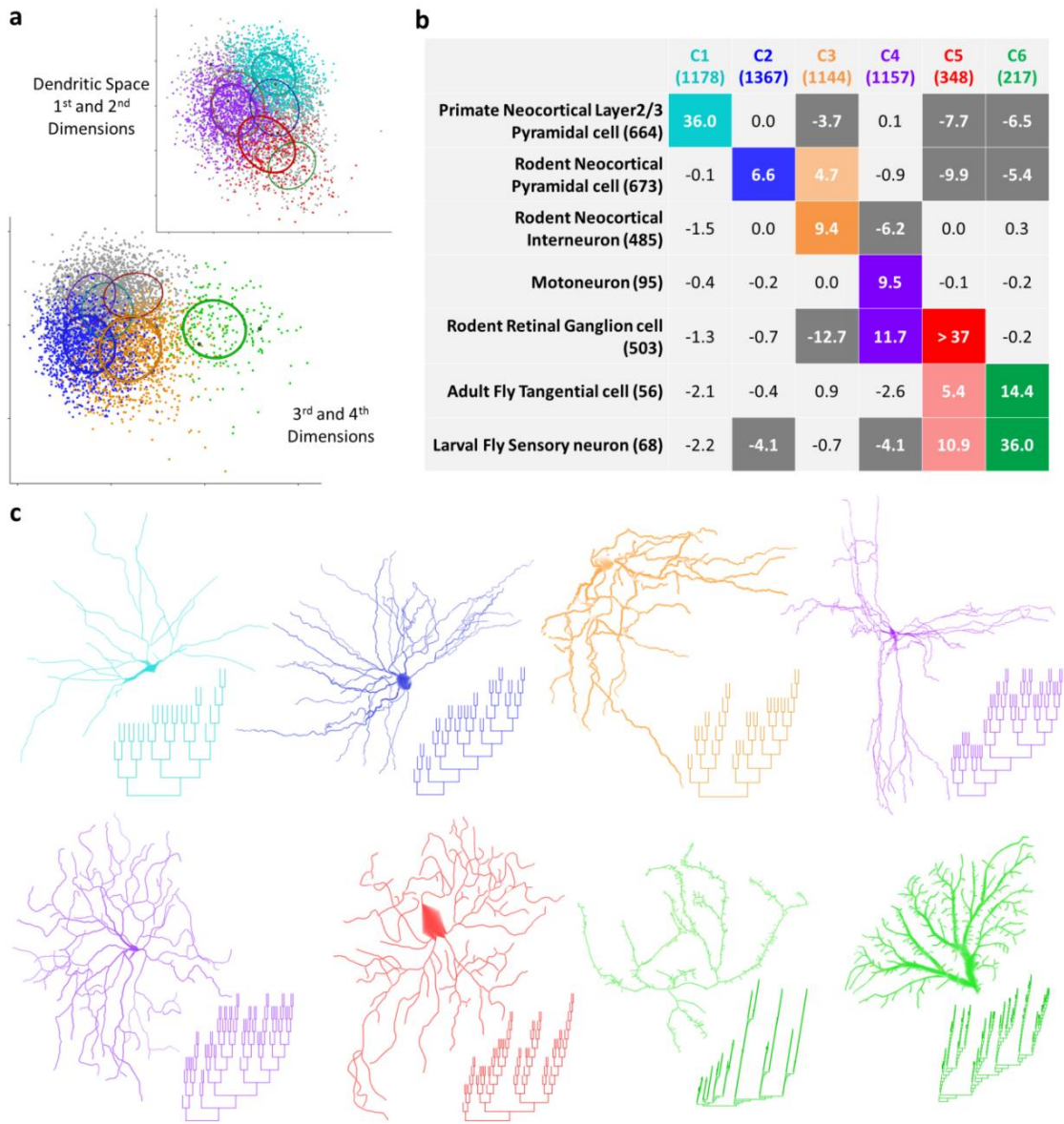


Figure 5.3. Dendrite clusters. **a.** Clusters of (non-apical) dendrites are shown in two perspectives of alignment space. Each perspective highlights dendrites in three of the six clusters, with dendrites in the other three shown in gray. **b.** The association matrix of the clusters and seven groups defined by metadata combination of species, region, cell type, and age range. The number of dendrites in each group and cluster are shown in parentheses. Values reflect the natural log of the contingency matrix p-values, with absolute values above 3 being significant ($p \leq 0.05$). Dark gray cells represent cases in which significantly fewer dendrites satisfy the cluster/group association than expected given the marginals. Positive values in colored cells represent cases in which more dendrites fall into the cell than expected. Darker colors signify that an example morphology and dendrogram is provided in **c** (NMO_05022, 09439, 00298, 00625, 05409, 06531, 07043, 06659 from [Cameron et al., 1991; Kong et al., 2005; Coombs et al., 2006; Cuntz et al., 2008; Anderson et al., 2009; Groh et al., 2010; Kubota et al., 2011; Sulkowski et al., 2011]).

Axons, though with far fewer total reconstructions ($N=1,230$) and representing fewer metadata groups, tend to be larger and so have the potential to separate more clearly. Perhaps due to their relatively limited diversity, axons exhibited structure in just two dimensions and fell into 4 clusters (**Figure 5.4a**). Nevertheless, the clusters (spherical but varied in size) had strong associations with metadata groups, specifically separating insects from mammals, pyramidal cells from different rodent species, and distinct interneuron types (**Figure 5.4b**). The fly olfactory neuron axons most clearly associated with cluster 1 and are substantially dissociated from clusters 3 and 4. Rat neocortical pyramidal axons associated with cluster 2, while mouse neocortical pyramidal axons associated with cluster 3. This result should be interpreted with caution since the larger axons of rat neurons may be more prone to slicing artifacts. Since the specific histological and imaging procedures can also affect the reconstructions, it is important to note that the mouse pyramidal axons primarily come from a single archive while the rat data have substantial contributions ($> 10\%$ each) from four different laboratories. Rodent (mouse and rat) neocortical dendritic targeting interneurons (primarily Martinotti and somatostatin-positive cells) also associated with cluster 3, though not as strongly. Rodent neocortical perisomatic targeting interneurons (primarily Basket and parvalbumin-positive cells) associated with cluster 4. Representative morphologies and dendrograms (**Figure 5.4c**) suggest features which are shared within or even define each cluster. Derivation of consensus representations and the qualitative and quantitative description of those features follow.

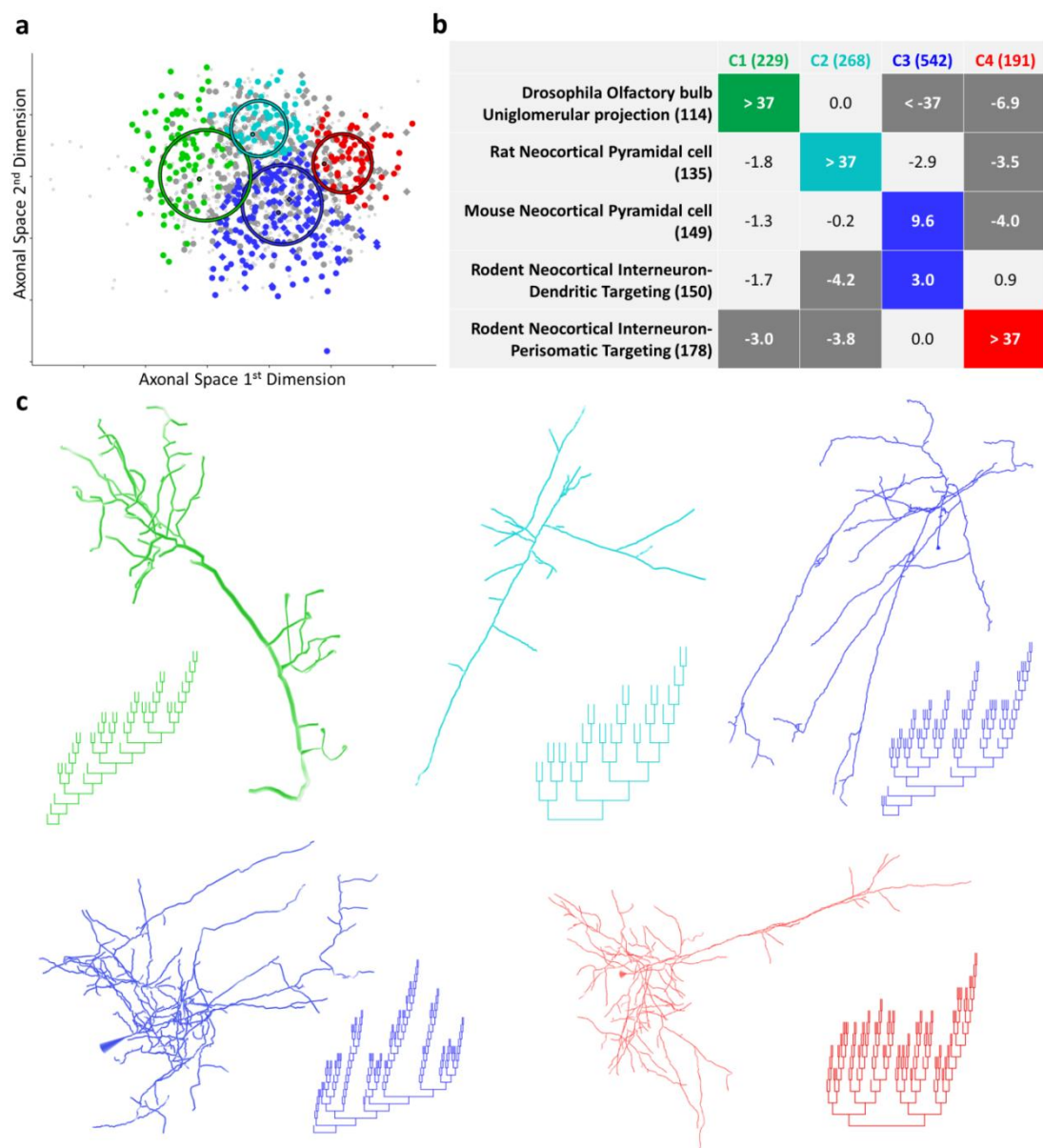


Figure 5.4. Axon clusters. **a.** Axons fall into four spherical clusters of varying sizes. Colored dots correspond to axons classified with the cluster associated with the matching metadata group. Diamonds represent rodent cortical dendritic-targeting interneurons. Circles are defined by the covariance matrix of their cluster model with a radius of one standard deviation. **b.** Association matrix for axon metadata groups and clusters. Colored and dark gray cells show significantly positive and negative associations, respectively, with the number of axons in each group and cluster in parentheses. **c.** Representative morphologies and dendrograms are shown for each significantly positive association (NMO_02574, 00315, 01209, 00424, 00424 from [Wang et al., 2002; Jefferis et al., 2007; Nikolenko et al., 2007]).

Multiple sequence alignment and consensuses

Multiple sequence alignment (MSA) provides the means for extracting common structural features of a group of neurite sequences and measuring their level of conservation. MSA can be visualized as vertical densities illustrating how represented various positions are (**Figure 5.5a**). A consensus structure can be defined by setting a threshold for the proportion of member sequences aligned at each position. A low threshold will result in a relatively larger consensus with lower average conservation (matching bifurcations) across member sequences. This choice tends to reveal common local patterns, though in the extreme, motif analysis (Gillette and Ascoli, 2015) could be more effective for that purpose. Higher thresholds, in contrast, reveal the common global pattern by producing narrower consensuses with higher average conservation. A 50% threshold was found to yield consistently clear results across clusters. In general, the relative consensus length (number of bifurcations normalized by the median sequence length in the group of neurites) and average conservation together provide a measure of how inter-related the sequences in a set are in terms of shared topological features. For a measure of global representativeness that accounts for the sequence length variability, we also employed the normalized pairwise alignment score.

We take for comparison axon clusters 3 and 4, which are next to each other in alignment space but display considerably different consensus features. In order to focus on the most representative trees of each cluster, we only considered those sequences within one standard deviation of their cluster centers. Furthermore, to avoid any potential effect of number of sequences on the MSA, the 120 sequences (the smallest extracted set size, from cluster 4) closest to their respective cluster centers were used for MSA and

subsequent analysis. The consensus of cluster 3 has an absolute length of 52, a relative length of 74.3%, and an average conservation of 75.8% (**Figure 5.5b**). The consensus of cluster 4 has a substantially longer absolute length of 117 but a smaller relative length of 67.8% due to the larger median sequence length relative to cluster 3 (**Figure 5.5d**). The cluster 4 consensus also has a higher average conservation of 79.6%, meaning that while cluster 3 has a higher proportion of well-aligned branches, the well-aligned branches of cluster 4 are more densely represented in the cluster sequences.

These properties are in part explained by the different variability of sequence length between the two clusters. If the higher variability of a cluster is due in part to a larger number of relatively short sequences, fewer bifurcations overall could align in the consensus. Longer sequences may be more likely to contribute to the consensus but will also align in many positions that do not pass the consensus threshold. Cluster 3 sequences have an average length of 139.7, median of 70.5, and standard deviation of 173.3; in contrast, cluster 4 sequences have an average length of 193, a median of 174, and a standard deviation of 98.8. The impact of this variability is seen in the branch lengths of the consensus dendrogram (**Figure 5b,d**). However, alignment scores normalized to z-scores (a score of 0 being equivalent to an average alignment between random sequences), show the same relationship with cluster 3 having a score of 1.42 ± 0.06 compared to the cluster 4 average of 2.77 ± 0.09 ($p = 6.7e-8$). This is consistent with the larger Gaussian model which was fit to cluster 3, indicating the looser relationship between its sequences. The consensus dendrogram of cluster 3 shows where a substantial portion of the variability falls. The root branch is particularly long, meaning that a

relatively large proportion of parts of member sequences exist prior to that bifurcation. Sample axons from the cluster illustrate the alignment of the axons and consensus in full context (**Figure 5.5c**).

Consensus structural details also differ substantially between clusters 3 and 4. Consensus 4 has greater symmetry (**Figure 5.5d**): in fact, consensus 3 has a maximum branch order of 13 compared to 12 for cluster 4, even though cluster 4 has 65 more bifurcations. The axons of cluster 3 show a greater average asymmetry of 0.611 compared to cluster 4 axons asymmetry of 0.54 ($p < 2.2e-16$). Other structural differences could relate to differential spatial targeting of these classes. Although the arbors from the two clusters showed similar spatial arrangements (measured by the distribution of weighted Euclidian distances of branches from the arbor center of mass), their total neurite length differed significantly. Cluster 3 had a median length of 10,050 μm (mean 14,530 μm) and cluster 4 a median of 12,952 μm (mean 19,450 μm ; Wilcoxon rank sum test; $p < e-13$). Although many of the larger neurons in the clusters were not from the associated cell classes, the trend was consistent. Pyramidal axons, with a median length of 3,261 μm , were shorter than dendritic targeting interneurons, with a median length of 5,440 μm ($p < 1.6e-15$). Perisomatic targeting interneurons, with a median of 6,671 μm were yet longer ($p < 2.0e-3$). The greater length in spite of similar spatial arrangement suggests a greater density or some alternative heterogeneous distribution of branches for cluster 4 axons.

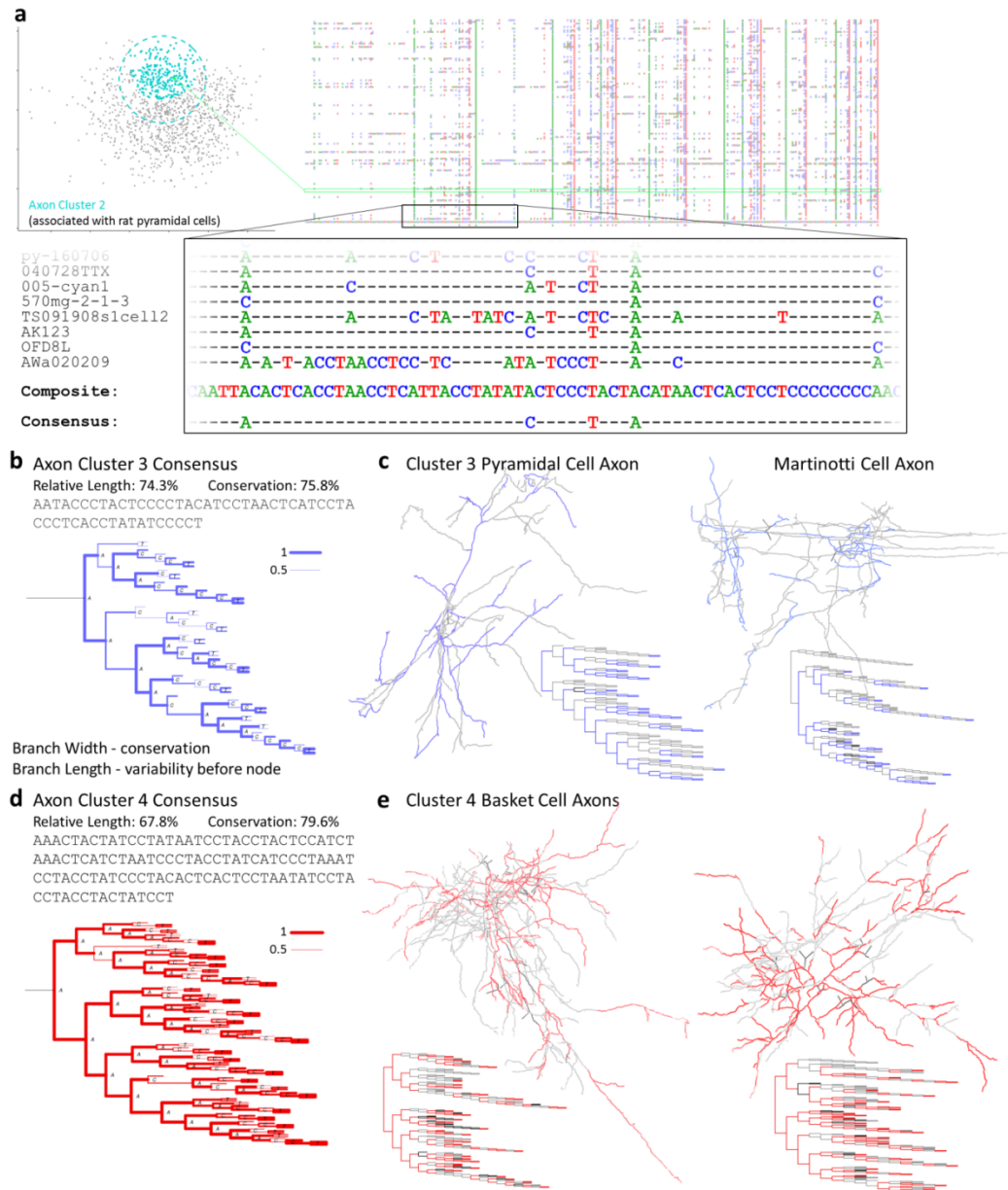


Figure 5.5. Multiple sequence alignment and consensuses. **a.** Sequences of axonal cluster 2 (left) are aligned (right). A section of the alignment is magnified along with the composite and consensus (bottom). A single neurite in the cluster and its sequence in the MSA is highlighted (green). **b,d.** Consensuses of axon clusters 3 (**b**, blue) and 4 (**d**, red) in sequence and dendrogram form, with relative length and average conservation statistics. Dendrogram branch width indicates conservation of the parent bifurcation (from 0.5 to 1). Branch length indicates variability preceding a bifurcation, calculated as the average proportion of a sequence falling between conserved bifurcations. **c,e.** Example morphologies and dendrograms of cluster 3 (**c**) and 4 (**e**) (NMO_02624, 00427, 07447, 00332 from [Wang et al., 2002; Ikegaya et al., 2005; Kawaguchi et al., 2006]). Colored segments indicate an alignment with the consensus at the bifurcation. Black and gray segments indicate consensus bifurcations not found in the neurite and vice versa.

Consensuses and motifs

We next considered whether the consensus is representative of local motifs patterns (Gillette and Ascoli, 2015), and whether motifs capture the global features of the consensus. Trimer-constrained surrogate sets were generated for each axon cluster consensus and for each individual axon sequence in order to produce motif profiles for comparison with the motif profiles of each cluster's sequences. Briefly, random trees were generated with just size, size and node-type proportions, and size and dimer proportions equivalent to each consensus. For a given consensus or axon sequence, node-type, dimer, and trimer proportions were then normalized to the respective proportions in the surrogate sets by determining the percentile rank of a given consensus k -mer proportion within the surrogate set's k -mer proportion distribution (see **Chapter Four**: for further details). The correlation between the motif patterns in the neurite sequences and their corresponding consensuses was small across the 4 axonal clusters (Pearson coefficient of 0.37), with the consensus values varying substantially more than the cluster averages (**Figure 5.6a**). Moreover, while clusters 1 and 4 had very similar motif profiles, with only minor differences in dimers AT, CC, AC, and CT, their consensuses differed dramatically (**Figure 5.6b**). For further comparison, cluster 3's values are very different from clusters 1 and 4, falling much closer to baseline, while the global consensus appears intermediate between those of clusters 1 and 4. Thus, while differences in global structure can in principle influence motif patterns, local structural features appear to be largely independent.

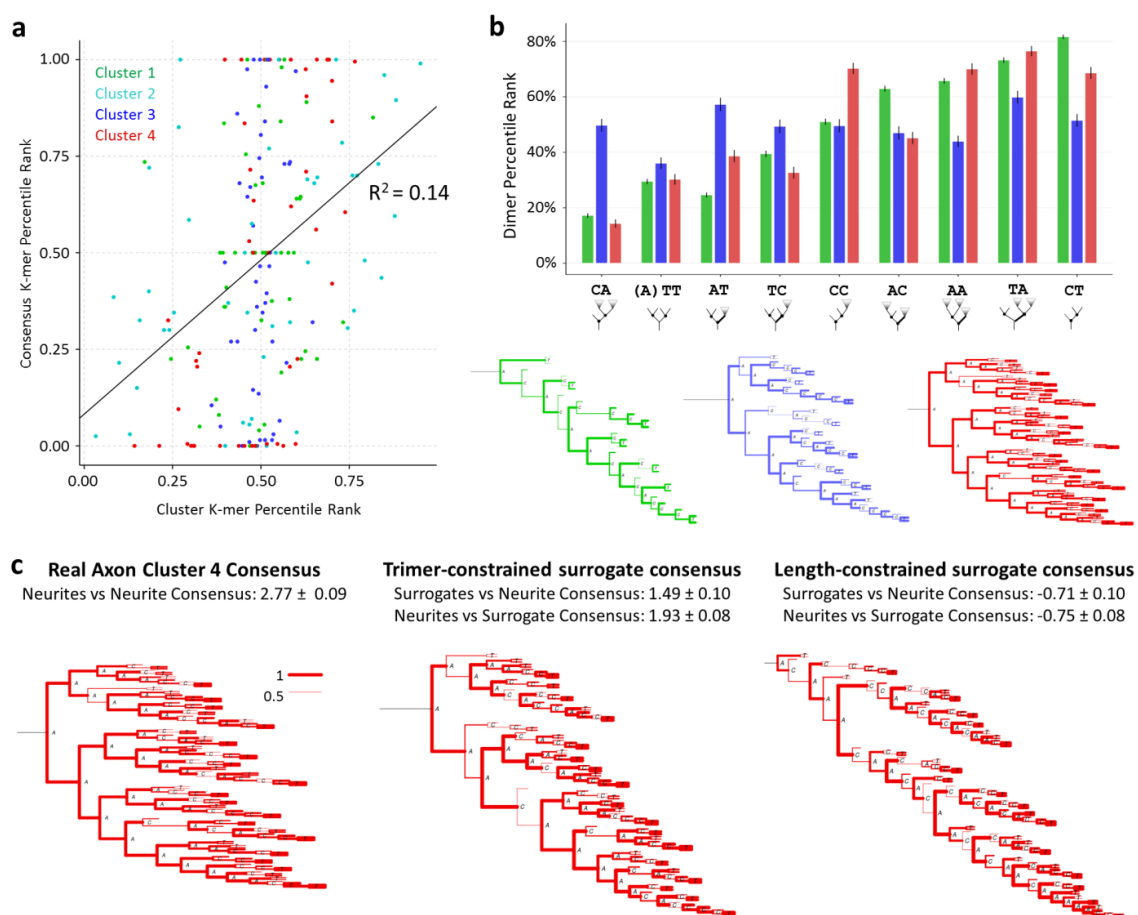


Figure 5.6. Global consensus and local motifs. **a.** Plot of the k -mer percentile ranks (normalized frequencies), for each dimer and trimer of each cluster, with the cluster's consensus values on the y-axis and the cluster's average sequence values on the x-axis. The correlation is small, showing that consensus does not capture well the local motif information about the cluster's members. **b.** Dimer profile (top) and consensus dendrograms (bottom) for clusters 1, 3, and 4. The dimer profile is similar between clusters 1 and 4, while the consensus differ dramatically. **c.** Consensus dendrograms of cluster 4 neurite sequences, trimer-constrained surrogates of the sequences, and length-constrained surrogates of the sequences. Average normalized alignment scores are provided given a set of sequences aligned to a single consensus. Dendrogram branch widths show parent bifurcation's conservation.

Motifs were similarly insufficient to consistently capture global features. Trimer-constrained surrogates of each cluster were aligned and consensus generated. Some trimer-constrained surrogate consensus were comparable to their real neurite consensus counterparts. For instance, global alignment scores of cluster 3 neurite sequences aligned

to the neurite consensus were not significantly different than the scores when aligned to the surrogate consensus. However, in other cases, most notably cluster 4, the trimer-constrained surrogates produced a far less representative consensus (**Figure 5.6c**). For comparison, length-constrained surrogates always failed to produce representative consensus with alignment scores systematically below those achieved with trimer-constrained surrogate consensus. This means that generally motifs and consensus have only limited impact on each other and they provide complementary information relating to local and global patterns, respectively.

Discussion

Biological and functional interpretations

Axons are of particular interest due to their role in connecting neurons to enable network function. Given the diversity of cortical inhibitory interneurons in terms of axonal morphology (DeFelipe *et al.*, 2013), it is not surprising that these cells would fall into different topological clusters. Perisomatic-targeting interneurons are responsible for generating synchronous network rhythms (Freund, 2003; Bartos *et al.*, 2007) by densely synapsing on a large number of principal neurons. The relative symmetry of perisomatic-targeting interneurons might then reflect a combination of dense synaptic fields along with growth mechanisms to ensure minimal path distance (balanced against wiring minimization) (Cuntz *et al.*, 2007). Shorter paths to targets would serve to optimize synchrony of the GABAergic signals from the interneuron.

It is interesting that a complementary subset of (dendritic-targeting) interneurons would instead associate with neocortical pyramidal cells based on the topological

sequences of the respective axonal arbors. This does not mean that the two axonal classes have the same distribution of topological features, as they do not share the cluster space equally; however their co-clustering does suggest a substantial overlap in those features. The relatively asymmetric branching of dendritic-targeting interneurons and mouse pyramidal cells suggests two possible non-exclusive explanations. The first is that the two classes are similar in their spatial targeting, and distinct from that of perisomatic-targeting interneurons. An alternative is that both classes are not as highly constrained in the timing of their efferent signals and thus wiring minimization is preferred over path minimization and its limiting effect on signal delay. The spatial targeting explanation is consistent with the roles of pyramidal cells and dendritic targeting interneurons in network computation requiring a diverse synaptic distribution to generate more powerful networks (Sporns *et al.*, 2004; Wang *et al.*, 2006). Specifically, sparser postsynaptic targeting would yield greater differences between the connectivity profiles of principal neurons and thus increase network complexity, which may aid in computational functions such as pattern separation (Knoblauch and Palm, 2001).

The differentiation between mouse and rat pyramidal cell axons is not as clear from a biological perspective. It is important to recognize, however, that most morphological data are derived from in vitro slice preparations, most often preventing the full reconstruct of axonal trees. Relative to dendrites, axons typically involve more expansive arbors and span well beyond the region explored in a given experiment. Given the size difference between mouse and rat brains, the impact of slicing on the reconstruction would be larger for rat axons. To test this hypothesis, we compared

artificially clipped mouse pyramidal axons to their rat counterparts. Our analysis suggests that morphological incompleteness due to sectioning artifacts is not the source of the difference between mouse and rat. Three pyramidal cell axons of various sizes (155, 495, and 1,725 bifurcations), were selected among the few instances of in vivo (putatively more complete) reconstructions, one from rat hippocampus and two from cat neocortex. To simulate the effect of tissue slicing, these arbors were modified by eliminating all branches and sub-trees beyond a given distance in the z-dimension from the soma. Cutting at several distances showed a decrease in normalized alignment score, but even with a quarter of the z-range and less than 50% of the original sequence length, the scores always remained significantly greater (four standard deviations above the mean) than length-matched random sequences. Other similarly severe cuts maintained much higher normalized scores. This suggests that sequence analysis is relatively robust to reconstruction artifacts of this kind. Thus, brain size is not likely to explain the differences between rat and mouse axonal trees. Other experimental protocol details, or an unknown biased sampling of neocortical pyramidal neurons, may yet be responsible; however, it is also possible that mouse and rat pyramidal cell axons indeed slightly differ in their branching topology.

Sequence alignment and expanded representations

The global sequence alignment modifications were necessary to ensure proper accounting of tree topology. Attempts to use standard global alignment (not respecting tree structure) and a simpler clustering method revealed no meaningful differences among known neuron types. This is because very different trees could have similar

sequences if the structural attributes of bifurcations are disregarded. For instance, three subsequences with distinct patterns, α , β , and γ , could be arranged in sequence as $\alpha\beta\gamma$ and matched to another sequence $\alpha'\beta'\gamma'$ where α' is a close match to α , β' to β , and γ' to γ . These two trees, however, could be arranged in very different ways: α as the parent of subtrees β and γ , and α' as the parent of β' and β' as the parent of γ' . Thus, $\alpha\beta\gamma$ and $\alpha'\beta'\gamma'$ would match well despite substantially different tree structures. In contrast, the modified global alignment algorithm implemented in PASTA distinguishes the two cases.

Expanded representations including branch-level morphological features could enhance alignment sensitivity, decreasing the likelihood of spurious alignment for shorter sequences. For instance, several characters could be used for each topological node type to encode the bifurcation angles (i.e. between child branches, between bisector and parent branch, or between child plane and parent) relative to the population distribution. The increased sequence complexity would alleviate the need for a gap opening cost, allowing a greater range for parameter tuning. Additionally, such encodings would enable analyses for relating local topological patterns with other morphological features and for better integrating morphological features into neuronal classification. Subtree size and tortuosity are similarly reliable branch metrics which could be encoded; distance from soma and maximum distance to termination are also suitable candidates. Diameter and associated metrics (e.g. surface area and volume), while possible in principle, are more susceptible to experimental protocols (Scorcioni *et al.*, 2004). Combining several morphological attributes in the encoding would multiply the number of characters required, but feature vectors could be used rather than characters, increasing sensitivity at

the expense of speed. Alternatively, the number of features and their discretization could be judiciously selected based on the specific research questions.

Leveraging pairwise tree alignment, cluster analysis, and multiple sequence alignment, an expanded encoding could help discover novel relationships between morphological features, including their relative conservation and specificity across neuronal classes. For instance, the different tortuosity among interneuron classes (Kawaguchi *et al.*, 2006) could reveal regional differences within the arbor, which would suggest axonal domains with specific targeting features. The level of conservation would provide an indication of the feature's functional importance as well as provide a constraint for modeling growth mechanisms.

Since morphology ultimately depends on ultrastructural features to produce function, lower-scale branch properties such as spines or bouton densities and even local molecular expression could create powerful analyses. This would require a richer data type and new experimental processes, but with recent technologies the prospect is real (Schmitz *et al.*, 2011; O'Rourke *et al.*, 2012; Deisseroth and Schnitzer, 2013). Combining biophysical data with pairwise alignment would enable exploration and detection of morphological-physiological relationships, such as spike failures near regions of high bifurcation density, or receptors of neurotrophic factors in sequences with many C nodes. The code available at <http://krasnow1.gmu.edu/cn3/NeuriteSequence/> includes sample expanded branch feature encoding to facilitate future exploration by the research community.

Optimizing algorithms and analysis

The sequence approach depends critically on the constraint that small-side subtrees necessarily be matched to other smaller-side subtrees. This condition produces near-optimal results (equivalent to the tree edit distance) with moderately to highly asymmetric trees, which have low probabilities of improved tree-alignment from a larger-side subtree aligned to a smaller-side subtree. The scores are near optimal even in the case of trees with greater symmetry, so long as the subtrees are topologically similar. This assumption is generally true of neurites, which explains its effectiveness in this study. However, it is possible to conceive a “worst-case scenario” of a root bifurcating into two large subtrees with exactly the same number of branches, but one fully symmetric and one fully asymmetric. Aligning the corresponding sequence with an almost identical arbor in which a single branch is added to the symmetric side (or a single C node removed from the asymmetric side) would yield a dismally low score despite the great structural similarity between the two trees.

To gauge the risk incidence of this effect in neurites, we assessed the asymmetry characteristics of bifurcations with similar subtree sizes, specifically, those with a difference of no more than 6 bifurcations. Of the 5% of axon, 26% of apical dendrite, and 36% of dendrite bifurcations satisfying the condition, the median difference in asymmetry was 0.14, with 75% having a difference less than 0.25 (95% with a difference less than 0.47), though the differences were much smaller for the 5% of subtrees with more than 50 bifurcations. This survey suggests that only dendrites might be susceptible to this problem, though with reasonably small impacts for each instance. Thus, based on

currently available data, the relatively simple topological encoding results in practically robust sequence alignments. However, the possibility that future larger dataset of complete axonal reconstructions will require a more optimal alignment method cannot be ruled out.

In the big-data prospect of orders-of-magnitude increase in neurite dataset size, heuristics will become important for efficient processing. Given that MDS does not require distances for every neurite pair, not all pairwise combinations are necessary for producing an alignment space. Moreover, for search or analysis, topological metrics can be used to limit the number of neurite alignments by filtering out those that would likely score poorly, as achieved by BlastNeuron with a broader range of morphometrics (Wan *et al.*, 2015). The PASTA algorithm itself can be sped up by ignoring the possibilities least likely to contribute to an optimal alignment (e.g. those corresponding to the upper right and lower left corners in the alignment matrix: see **Figure A1.1c**). Other heuristic approaches could be applied that involve candidate seed locations of short matching strings from which the alignment is extended in either direction (Altschul *et al.*, 1997). The most efficient modern algorithms index seeds in advance (Buchfink *et al.*, 2014). A larger dataset of more complete axons, widely expected from ongoing and discussed connectomic projects, would require appropriate heuristics and efficient implementation in a compiled language, such as C or C++. In cases where sensitivity is more important than speed, sequences of branch feature vectors could take the place of discrete characters.

The diversity of available gene sequence alignment tools highlights the utility of specific parameters and alternative algorithms for different experimental questions or computational requirements (Ruffalo *et al.*, 2011; Thompson *et al.*, 2011; Berger *et al.*, 2013; Buchfink *et al.*, 2014). Methods of neurite pairwise alignment of one sort or another, including variations on our method, will likely grow in a similar manner to fill various niches, whether in terms of sensitivity or efficiency.

Conclusion

The newly developed method for aligning binary tree sequences enabled the comparison of large numbers of neurites and revealed their shared topological features. Sequence alignment similarity with a purely topological encoding was sufficient to distinguish different arbor types (axons, dendrites, and apical dendrites) as well as certain neuron types, brain regions, and species. Moreover, the sequence representations could differentiate cell classes better than (or at par with) traditional morphometrics and are consistently as informative as traditional topological metrics. Cluster analysis produced groupings highly associated with known metadata differences, and multiple sequence alignment generated a consensus representation of each cluster that revealed the common topological features. Comparative analysis between consensus and motif profile demonstrated that these analyses captured complementary topological characteristics, with consensus reliably extracting global features and motifs effectively quantifying local features.

Acknowledgements

We thank Duncan E. Donohue, John J. Grefenstette, Matteo Mainetti, and Iosif I. Vaisman for insightful discussions and ideas. This work was supported by NIH grant R01 NS39600.

APPENDIX ONE: DETAILS OF GLOBAL AND MULTIPLE ALIGNMENT PROCESSES

Global sequence alignment

Global alignment of tree-sequences uses a modified Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) in which a matrix is created from the character positions of the two sequences being aligned. Scores are calculated from the beginning of both sequences (top-left) to the end of both sequences (bottom-right) using previously calculated position alignments. The best score from the three possible steps to a given position is assigned to each position in the matrix (**Figure A1.1a**). Those steps include a match between the characters (diagonal step from above and to the left), a gap over the character of the first sequence (a step from above), or a gap over the character of the second sequence (a step from the left). The alignment score is given in the final position and a back-trace of the optimal route from the final position to the start position produces the alignment (**Figure A1.1b**).

The tree-based sequence modifications correspond to the A, C, and T bifurcation node types, allowing for a representation that includes branch properties in addition to local topology. It also assumes each sequence is a single complete tree. Limitations are placed both on matching and gapping based on node type in order to ensure the alignment is consistent with a valid tree structure (**Figure A1.1a**). Only pairs of characters of the same type are directly matched in the dynamic programming process. A-type characters

cannot be explicitly gapped. This can result in inaccessible positions during back-trace, however a valid back-trace will exist as long each sequence is a complete tree. C-type characters can be gapped normally. T-type characters can be gapped, but the gap must extend until the T node's associated A node. In terms of tree structure, the T node would be the last traversed node on the A node's smaller subtree (assuming *StL* traversal). This within sequence A-T relationship is defined in a recursive manner such that a T node's associated A node is the nearest prior A node that does not have its own nearer subsequent T node. While an entire subtree can be eliminated by gapping all characters from A to T, it is also possible for all but a single branch to be eliminated, leaving a single C node. Thus, when gapping a T node and its associated subtree, if the character in the opposing sequence position is a C, the gap region can end such that the A and C are matched. The final T node of any sequence will have no matching A and cannot be gapped. This would effectively lead all tree-sequences to align on the final T node making certain alignments impossible, such as a smaller tree aligning to an inner subtree of a larger tree. Two final modifications fix this problem. First, A nodes can be gapped prior to any matches. Second, the final T node of either sequence can gap back to another T node that has a match.

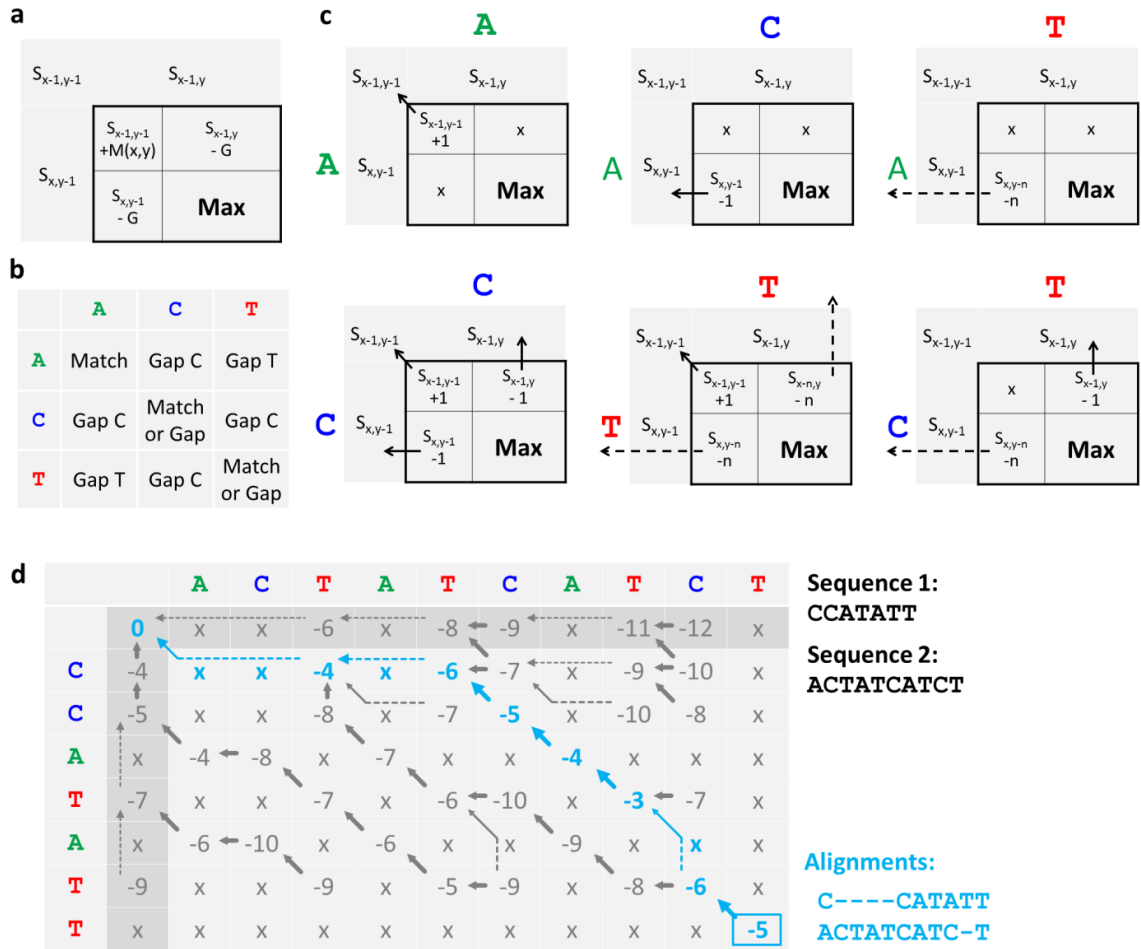


Figure A1.1. Dynamic programming alignment of tree node types. **a.** The first table shows the generic calculation of scores for a given matrix position (x,y) using the scores of cells to the left, top, and top-left. The score from the top-left is added to the match score for the characters at (x,y). The score for gapping from above equals the score above (x-1,y) minus the gap penalty, while the score for gapping from the left is that score (x,y-1) minus the gap penalty. The cell's score is then the maximum value of the match and gap scores. **b.** Cells with certain node type pairs have restrictions as to whether a match, left-gap, or right-gap is allowed. **c.** Each node type pair is shown with arrows pointing to where the scores are coming from. Dashed arrows for gapping a T node represent a multiple position gap based on the requirement that the gap continue until the start of the node's subtree. **d.** Complete alignment matrix for the sequences shown. The blue path represents the optimal back-trace. Each cell with a score has a backward arrow leading back to the preceding cell that produced the optimal score. Arrows with dashes followed by a diagonal solid line represent an alignment of a C node with a larger subtree, such that the gap begins at a T and back-traces to an A that matches the C.

Score baseline and distance normalization

Specifically, each alignment score was normalized to a baseline of alignment scores given random trees of the same sizes of the neurite sequence pair. For each sampled length, 1,000 tree shapes were generated randomly and encoded as sequences. Sequence alignment was performed between ordered sequences for each length pair such that the n th sequence of length L_1 was aligned with the n th sequence of length L_2 to avoid any dependence between scores, resulting in 1,000 scores for each length pair. Each score was then converted to a per-character score (S_{pc}) based on the length of the shorter sequence. This was achieved by adding back to the score the length difference of the two sequences, as well as the value of one gap opening (G), and dividing by the shorter sequence length. An optimal per-character score would be 1 and the lower bound for the score is limited only by the size of the sequences and the potential number of gaps and gap openings. With no gap open cost the minimum per character score would be -1.

$$S_{pc} = \frac{S + |L_1 - L_2| + G}{\min(L_1, L_2)}$$

Equation A1.1

The mean and standard deviation of those per-character scores was taken for each length pair. Means and standard deviations of length pairs falling between those sampled were interpolated. All real neurite per-character scores were converted to z-scores based on the per-character score mean and standard deviation of the neurite pair's lengths (**Figure A1.2**). Each z-score was converted back into a per-character score using a set mean and standard deviation, creating a fully normalized score (S_{norm}). The mean and standard deviation were chosen such that the probability of a perfect per-character score

of 1 would be exceptionally low. Finally, the fully normalized scores were converted to distances by subtracting each score from the optimal score of 1. In the rare case of a fully normalized score above 1, resulting in a distance value less than 0, an exponential function was applied making the distance greater than 0. Baseline score normalization was calculated in R prior to further analysis.

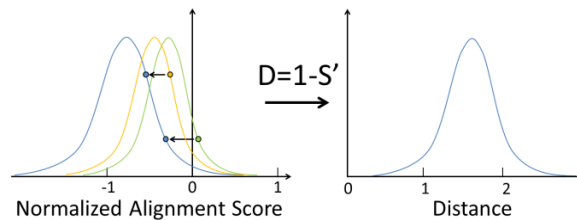


Figure A1.2. Normalizing the per-character alignment score. A given per-character alignment score has a z-score with respect to the distribution of baseline scores for the specific length pair of the aligned sequences. The yellow and green points represent scores from alignments with different length pairs. The lines represent the baseline distribution of per-character scores. Each point is then normalized (arrows) using their respective distribution to the mapped value along an arbitrary length pair distribution (blue points). The distance values are then calculated by subtracting the normalized per-character scores from 1, the maximum possible score.

Additional methodological details

Non-metric MDS

Non-metric MDS is performed by embedding by some process (usually an attempted standard MDS algorithm) then iteratively shifting element location until *stress* is minimized. *Stress* is the measure of how different each neurite pair's position is in the original ordered list of distances and in the ordered list derived from the embedded space. Thus, the two neurites with the smallest distance between them in the original data should also be the two nearest neurites in the new space, while two neurites that were an average distance apart in the original data will be about an average distance apart in the new space relative to other neurite pairs.

The *MASS* (Venables and Ripley, 2002) R package was used to perform non-metric MDS on datasets consisting of all neurites and separate instances of the data composed of neurites from each arbor type. In each case a sufficient number of dimensions were used such that the stress was at or below 15% and was decreasing by less than 1% with each additional dimension. Inspection of the spaces showed no effective differences in the primary dimensions used for analysis between spaces generated using more dimensions. Moreover, the spaces produced for each arbor type were very consistent with the space produced for all arbors.

Model-based clustering

We used the *mclust* (Fraley and Raftery, 2002; Fraley *et al.*, 2012) R package for model-based clustering, which performs Expectation Maximization (EM; Dempster et al., 1977) to generate optimal parameter values to fit the data to a set of multivariate Gaussian models. Variable Gaussian features include size, shape (different size in each dimension), and orientation. Shape and orientation may be unspecified, resulting in spherical clusters (similar to k-means clustering), or ellipsoidal clusters without orientation. Each parameter type could be equivalent or variable across clusters. These options are not exhaustive, but for the purposes of exploration they are effective and provide a useful limit on an otherwise unlimited set of models. Multiple parameter sets and number of clusters were tried and a Bayesian Information Criterion (BIC) value produced for each. The BIC is the measure of the log likelihood of the model given the data minus the complexity of the model (in terms of number of parameters) times the log of the dataset size.

$$BIC = \ln(L) - k * \ln(n)$$

Equation A1.2

Classification

In order to maximize classification performance and determine the most informative features, clustering was run on all possible permutations of alignment space dimensions and all possible permutations of topological metric principal components for a specified pair of neurite groups. Models were limited to 2 clusters, and for alignment space Gaussians were limited to equal size and no orientation.

LDA was performed by the R *MASS* package (Venables and Ripley, 2002), while feature selection for LDA was performed using R package *klaR* (Weihs *et al.*, 2005). Variables could be added or removed in the variable selection process and the improvement threshold for updating the model was a 1% improvement in accuracy. A 10-fold cross-validation was used, leaving 10% of the data for testing the accuracy of the final model. Since *stepclass* is stochastic with regards to the division of the data, different models were generated with different accuracy values in multiple runs. The process was run 10 times for each case and the mean accuracy was taken. The optimal models were found most times and always provided the final accuracy value. The order of mean, median, and max accuracy scores between variable sets (alignment space and topological metric PCs) were consistent.

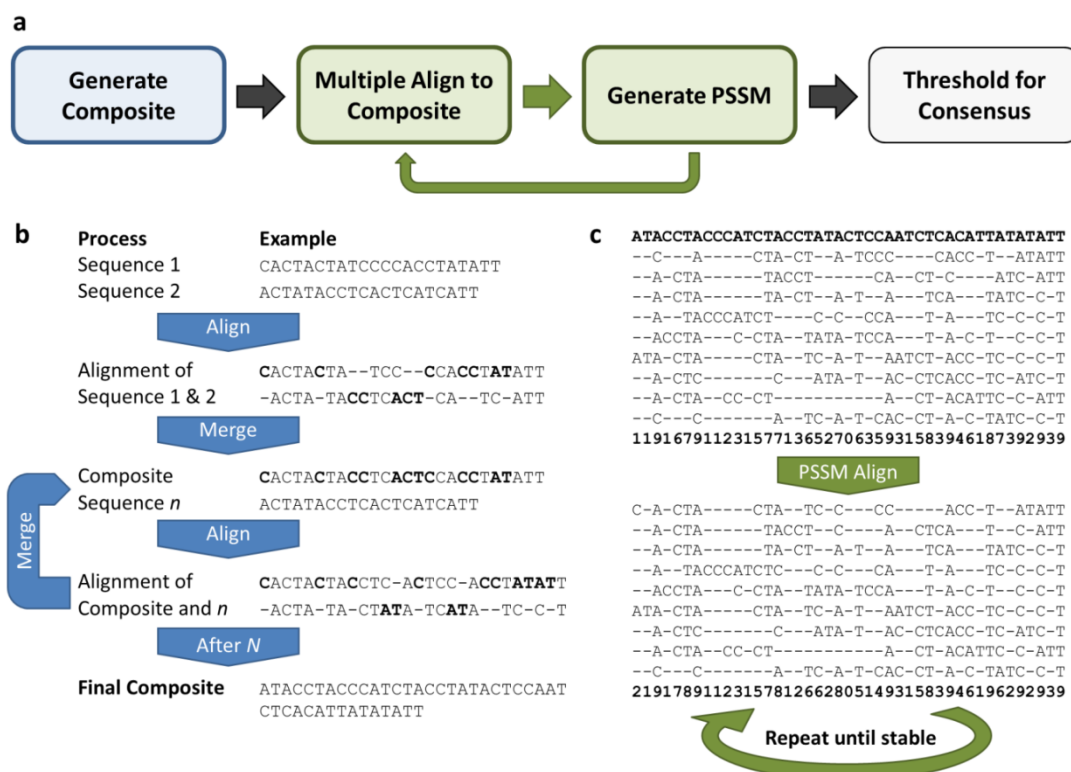


Figure A1.3. Multiple sequence alignment algorithm. **a.** Flow chart of the multiple alignment process, ending in a multiple alignment and a thresholded consensus. **b.** The process of building a composite begins with the matching of two sequences from the set to be aligned. The alignment of those sequences is merged into the composite which is then aligned to the next sequence. The process repeats until all *N* sequences have been merged into the composite, forming the final composite. **c.** All sequences are aligned to the final composite, creating a multiple alignment. A position-specific scoring matrix (PSSM) is generated based on the conservation of each position (i.e. how many sequences align at the position). The sequences are again aligned to the composite, but this time the score of a match is determined by the PSSM. The process repeats until the multiple sequence alignment is stable.

APPENDIX TWO: QUANTIFYING NEURONAL SIZE: SUMMING UP TREES AND SPLITTING THE BRANCH DIFFERENCE

[Todd A. Gillette*, Kerry M. Brown*, Giorgio A. Ascoli: in *Seminars in Cell & Developmental Biology* (2008) 19:485-493]

[Used with permission from the publisher; * These authors contributed equally]

Abstract

Neurons vary greatly in size, shape, and complexity depending on their underlying function. Overall size of neuronal trees affects connectivity, area of influence, and other biophysical properties. Relative distributions of neuronal extent, such as the difference between subtrees at branch points, are also critically related to function and activity. This review covers neuromorphological research that analyzes shape and size to elucidate their functional role for different neuron types. We also introduce a novel morphometric, "caulescence", capturing the extent to which trees exhibit a main path. Neuronal tree types differ vastly in caulescence, suggesting potential neurocomputational correlates of this property.

Introduction

Neurons play vital roles in central and peripheral nervous systems, including integration, production, and transmission of electrochemical signals, and establishment of plastic network connectivity. From prenatal development through adulthood, these functions are determined by many interacting factors, such as gene expression, intracellular molecular dynamics, and the type and distribution of ion channels. This

review focuses on the most visible, though quite complex traits of neuronal morphology. Morphological properties include a variety of general size characteristics such as length and volume, as well as attributes associated with complexity such as number of branches and asymmetry (Uylings and van Pelt, 2002; Scorcioni *et al.*, 2008). Each feature has implications for how neurons grow and adapt (Cline, 2001; Wong and Ghosh, 2002), as well as function both locally and in the broader neuronal network (Ascoli, 1999; Jan and Jan, 2003).

Neuromorphology

Neurons display elongated branching structures stemming from their cell bodies. These trees, called axons and dendrites, respectively send and receive signals to and from other neurons. Axonal and dendritic arbors differ in size and shape both between and within cell classes (**Figure A2.1**). Given the complexity and variability of neuronal morphology, computer-aided methods are typically employed to perform comprehensive and quantitative analyses. Many of the previous and current results reviewed here are based on 3D digital reconstructions of neuronal arbors (Ascoli, 2006). This process begins with histological slices laid out on slides. The neurons of interest are labeled so they can be captured using a digital camera mounted on a microscope, allowing focusing for acquisition of an entire 3D structure. Specialized software programs, such as Neurolucida (MicroBrightField, Inc., Colchester, VT), Neuron_Morpho (Brown *et al.*, 2005) (<http://www.personal.soton.ac.uk/dales/morpho>), or Neuromantic (<http://www.rdg.ac.uk/neuromantic>) provide tools for visualizing the captured images and/or tracing neuronal branches, outputting 3D coordinates and connectivity of trace

points. The end product is a digitized, morphologically realistic neuronal representation that can be loaded into software programs specialized in extracting an extensive array of morphological metrics, such as L-Measure (Scorcioni *et al.*, 2008) (<http://krasnow1.gmu.edu/cn3>). Digital reconstructions are further imported by software that simulate biophysical properties (Bower and Beeman, 2003; Hines *et al.*, 2004), thus providing morphological realism for electrophysiological modeling.

Neuron size and shape

Overall size encompasses an important subset of morphological characteristics of neuronal arbors. Distinct functional aspects of internal neuron size are defined by the total number of terminal branches (the “degree” of the tree) or by continuous morphometrics including total length, surface area, and internal volume (Ascoli *et al.*, 2008). The spatial extent of the neuron can also be described in terms of distance reached from the soma or by the height, width, and depth of a box containing the whole arborization. Greater length often corresponds to greater area of invaded space, and thus greater potential connectivity. For axons this may result in increased divergence (one signal sent to many cells). Total wiring is minimized by increasing length for axons over dendrites (Chklovskii, 2000) when divergence is higher than convergence (more presynaptic than postsynaptic neurons). When convergence is greater than divergence, dendrites have relatively greater length. These predictions were confirmed in retinal, cerebellar, olfactory bulb, and neocortical neurons.

Membrane surface area correlates with number of synapses in dendrites and axon terminals (Pierce and Milner, 2001; Liu, 2004). Signal propagation speed linearly

increases with diameter, however the metabolic cost also escalates due to the squared relationship between diameter and cross-sectional area (Wang *et al.*, 2008). Within the generally large metabolic costs of the brain, synaptic transmission and spike generation are particularly energy intensive due to the high ATP requirements of ionic pumps (Laughlin *et al.*, 1998). The total number of branches (defined as the regions between two bifurcations or between a bifurcation and a termination) is another measure of absolute tree size and varies widely between cell types (**Figure A2.1b**). Individual branches may operate in isolation for the separate processing of specific groups of synaptic inputs (Losonczy *et al.*, 2008). Bifurcations also serve as nodes of integration, enhancing dendritic computational power (Mel, 1994).

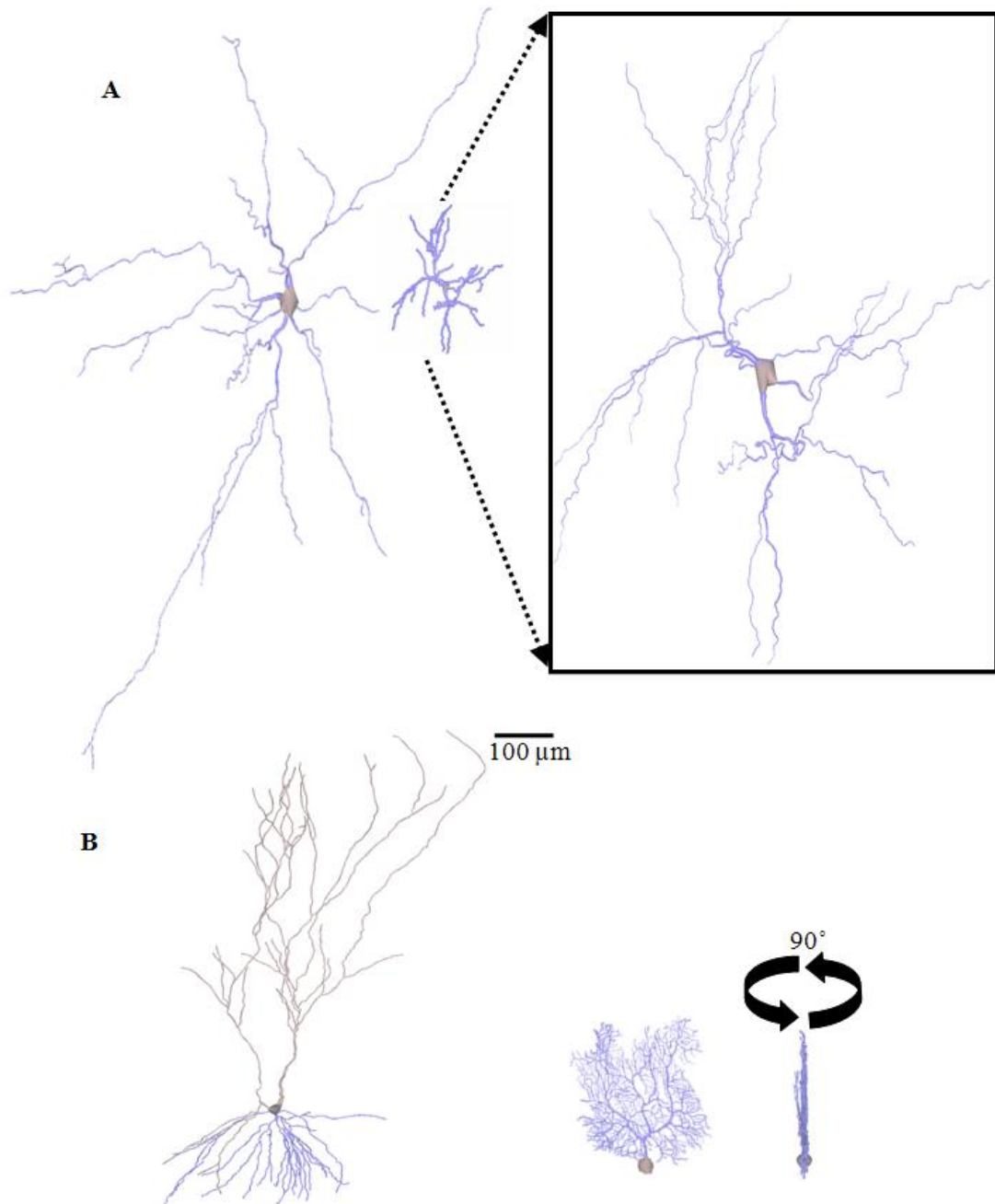


Figure A2.1. Diversity in neuronal size. **A.** Brainstem motoneuron (Núñez-Abades *et al.*, 1994) (left) and cerebrocortical stellate dendritic trees (Staiger *et al.*, 2004) (right) have similar shapes and numbers of branches (53 and 51, respectively) but different total lengths (6298 to 1966 μm). **Boxed Inset:** Zoomed-in (4x) stellate cell emphasizing morphological similarity to motoneuron besides size. **B.** Hippocampal CA3 pyramidal (Henze *et al.*, 1996) (left) and cerebellar Purkinje cell (Vetter *et al.*, 2001) (right) dendritic trees have similar lengths (9670 and 9883 μm) but different numbers of branches (121 and 567). Rotation shows the prototypical Purkinje cell planar shape. All trees (from NeuroMorpho.Org) are to scale.

Each path within a tree, from soma to termination, has its own total number of branches, length, surface area, and volume. For any size metric, the maximum extent among all paths in a tree is also functionally relevant. Greater distance along the path augments passive signal attenuation (i.e. when voltage gated channels are not actively propagating the membrane potential), and thus decreases the influence of distal synapses (Pyapali *et al.*, 1998). Branch order, defined as the number of bifurcations from the cell body (**Figure A2.2a**), also influences signals. Computational models and electrophysiological experiments have shown that branch points are prone to failure in active signal propagation (Spruston *et al.*, 1995; Migliore, 1996; Williams and Stuart, 2000). These failures are due to impedance mismatch between branches, which is in turn affected by differential section areas (Migliore *et al.*, 2005). A greater density of bifurcations therefore can have a significant impact on passive and active signal propagation both towards and away from the soma. With active propagation, the density of active channels can modulate the effects of morphology and prevent the failure of dendritic spike forward- or back-propagation (Vetter *et al.*, 2001). Thus, branch points and length can respectively act as digital filters for active signals and continuous filters for passive signals.

Cell growth in development

All the aforementioned morphological properties mature through various developmental stages. Initially, neurons stem several small, undistinguished branches from the cell body. After multiple phases of retraction and elongation, one of the branches becomes differentiated by rapidly elongating and adopting axonal

characteristics (Dotti *et al.*, 1988). The rest of the sprouts eventually assume dendritic properties such as forming spines, small protrusions that synapse with excitatory axons (Scott and Luo, 2001). Axons and dendrites also develop distinct biophysical properties including channel distributions and cytoskeletal makeup (Fukata *et al.*, 2002). There are numerous molecular agents involved in axonal (Yoshimura *et al.*, 2006) and dendritic outgrowth (Scott and Luo, 2001), though their details are beyond the scope of this review (see also Donohue and Ascoli, 2005).

Morphological plasticity

Neurons arborize either from their growth cones located at extending terminal tips or from intermediate segments of existing branches (“interstitial” branching) (Dailey and Smith, 1996; Szebenyi *et al.*, 1998). Branch elongation and arborization allows synapsing with other cell processes, creating a neuronal network. The connectivity of the network must be refined through retraction of underused synapses, testing of others, and stabilization of functionally meaningful ones (Luo and O’Leary, 2005). This plasticity is driven by dynamic factors, particularly afferent activity, that can impact neurons throughout development and in different ways depending on age and cell type (McAllister, 2000; Cline, 2001; Wong and Ghosh, 2002; Jan and Jan, 2003). Dendritic growth mechanisms are sensitive at the population level, producing larger arbors in organisms with enriched environments (Fiala *et al.*, 1978; Nilsson *et al.*, 1999) and smaller trees and brain regions in response to stressful environments (Magariños *et al.*, 1996; Bremner *et al.*, 1997; Radley *et al.*, 2004).

Afferent activity also plays a role in retraction due to competition between inputs. Mitral cells in the olfactory system initially grow dendrites that contact multiple glomeruli, only later to retract from all but one (Malun and Brunjes, 1996). Multiple cerebellar climbing fiber axons synapse onto individual Purkinje cell dendrites. Upon innervation by parallel fibers onto the Purkinje cell, all climbing fibers retract, except a single one which then expands its dendritic influence (Scelfo and Strata, 2005).

Extrinsic vs. intrinsic influences

These steps in dendritic and axonal development temporally overlap and interact dynamically. This complex process is mediated by a number of intrinsic (host-cell specific) and extrinsic (cues from neighboring cells, target axons, and surrounding extracellular milieu) factors. The mutual interactions make it difficult to distinguish individual factor roles in dendritic and axonal growth. Models and experiments have each reached differing conclusions on the relative extent of intrinsic and extrinsic influence on neuronal growth.

Samsonovich and Ascoli (2003) found that the direction of growth for reconstructed hippocampal dendrites was mostly determined by repulsion from their own soma. Deviations from the orientation leading directly away from the soma were typically followed by corrective deflections based on the segment's resulting position. The repulsion strength decreased with distance from the soma, yielding the approximately conical shape of hippocampal principal cells.

A stochastic computational growth model could reproduce several morphologically diverse cerebral and cerebellar cortical cell types by modifying only two external cues (Luczak, 2006). Modifying the spatial distribution of simulated ‘neurotrophic particles’ changed both orientation and branch density, whereas modifying the dimensions of the space available for growth changed the gross size and shape of the cell. This study also suggested the requirement of an intrinsic constraint on maximum size to keep larger cells from extending at faster rates than neighboring neurons. Without such a correction, bulkier arbors would increasingly restrict growth particles from reaching smaller trees. The author theorized an alternative extracellular constraint based on distance-dependent concentrations of growth-facilitating cues, possibly related to cortical layers.

Qualitative and quantitative neuromorphological features

Tree shape is often the most visually obvious identifier of neuronal arbor and cell types. For example, pyramidal cell dendrites are composed of two polarized arbors, apical and basal (**Figure A2.1b**). Both grow away from the cell body in opposite directions, receiving input from different sources. Basal dendrites spread out as they extend from the soma, taking a conical shape with the apex located at the cell body (Henze *et al.*, 1996). Apical dendrites also spread out as they grow away from the soma, but they elongate more than basal dendrites, differentiating further as they advance through multiple layers.

Pyramidal cells are the principal neurons in both the cerebral cortex and hippocampus. Although all pyramidal cells share a relatively similar appearance, some

differences also exist between regions. The overall size of adult male rat pyramidal dendritic arbors (labeled *in vitro*) from the NeuroMorpho.Org (<http://NeuroMorpho.Org>) database (Ascoli, 2006; Ascoli *et al.*, 2007), measured as total number of branches, length, surface area or volume, is approximately 4 to 9 times larger for hippocampal than for cortical pyramidal apical arbors (detailed results not shown) and approximately 3 times larger for hippocampal basal arbors ($p < 0.001$ for all metrics). There is less difference between cell types when measuring pyramidal dendrites along their principal axis of growth, where cortical pyramidal dendritic arbors (apical and basal combined) are ~70% the size of hippocampal pyramidal dendritic arbors ($p < 0.05$). Hippocampal neurons are relatively spread out along the other two orthogonal axes compared to cortical pyramidal cells, which appear more elongated. Cortical dendrites may be confined within a more restricted lamella thus appearing flatter. Differences in tissue slice shrinkage leading to greater distortion in one of the cell types may also explain the results.

Cerebellar Purkinje cell dendrites are mostly restricted to growth in two dimensions. They fan out significantly in the rostrocaudal direction, but are nearly flat mediolaterally (**Figure A2.1b**, right). Purkinje cells are thus aligned in neat rows correlating with the deep folds (folia) throughout the cerebellar cortex. This architecture may help minimize the length of shared parallel fiber inputs (Eccles *et al.*, 1967). Some dendritic types, such as in retinal ganglion cells, stop growing upon reaching neighbors of the same type. The resulting ‘retinal tiling’ spatial layout (Amthor and Oyster, 1995) suggests repulsion between neighboring neurons as a shape-constraining mechanism.

Other neuronal types such as stellate cells and motoneurons grow away from the soma in many directions (i.e. they are multi-polar) (**Figure A2.1a**).

Underlying the shape and physical qualities of dendrites and axons is the cytoskeleton. The cytoskeleton provides and maintains the structural properties of the neuronal trees, mediating intracellular transport (Bhaskar *et al.*, 2007), and branch diameter (Hillman, 1979, 1988), elongation, and bifurcation (Dent *et al.*, 2003). Different cytoskeletal fibers include actin filaments, microtubules, and other intermediate filaments. In dendrites, actin is largely found near the branch surface, where it can give rise to high densities of spines, as well as forming filopodia that sample the extracellular environment to guide direction and growth patterns determining overall shape (Fischer *et al.*, 1998). Microtubules make up much of the branch core and act as a skeletal backbone maintaining the neuronal tree shape (Scott and Luo, 2001). Microtubules have polarity which determines the direction of resource transport and is an important differentiating factor between axons and dendrites. Axon microtubules only transport distally (towards the plus-ends), while dendritic microtubules face both directions (Baas *et al.*, 1988). The loss of function of a motor protein that drags the microtubules into dendrites minus-end first results in elongated dendrites without taper and with an organelle makeup closer to an axon (Yu *et al.*, 2000).

The size metrics discussed so far relate to many neuronal functions. However, other functional properties depend on more complex shape and size metrics such as branching distributions and patterns (Schaefer *et al.*, 2003). Branching patterns and distributions can be measured in numerous ways providing a challenging choice for the

best metrics to explore for a given cell type and features of interest. The following section reviews some common topological measurements including partition asymmetry and modifications thereof.

Beyond size: complexity and asymmetry

Considering the shape of neurons beyond their size opens the door to a more extensive set of metrics. Early measures of complexity were restricted to 2D images of neurons. Sholl analysis is a common metric that counts the number of branches at regular distances from the soma (Sholl, 1953). Modified (or Sholl-like) analyses, including 3D variations, were natural extensions once digital 3D reconstructions became available (Caserta *et al.*, 1995; Ristanović *et al.*, 2006). Sholl-like analyses expand size metrics by analyzing their distribution relative to distance measures such as branch order or path distance. As an example, dendritic branches tend to thin as they elongate. Accordingly, numerous computational modeling studies have used distance-dependent diameters to stochastically produce realistic neuron/arbor types (Burke *et al.*, 1992; Ascoli, 1999; Donohue and Ascoli, 2005b; Samsonovich and Ascoli, 2005).

Krichmar *et al.* (2002), loading CA3 pyramidal cell reconstructions into an electrophysiological simulation environment, found that size morphometrics affect firing activity but cannot account for all aspects of electrophysiological behavior. How dendrites are distributed across branch order significantly correlates with aspects of firing patterns. These results suggest that complex morphological properties must be characterized not only by overall size, but how size varies within the given tree.

Partition asymmetry

Morphological asymmetry reflects the relative balance of branches or size within a tree. Partition asymmetry (A_p) is a relative measure of asymmetry at a bifurcation node based on the distribution of degree (number of terminals) between the node's two subtrees (Van Pelt *et al.*, 1992). **Figure A2.2b** presents the formula and calculation for a sample tree. The value ranges from 0 (symmetric) to 1 (asymmetric). Since degrees are integers, the number of possible values of A_p depends on the node total degree. The denominator represents the degrees of freedom (total degree minus two, as both left and right subtrees must have at least one terminal).

The number of possible tree shapes increases exponentially with tree size. Even relatively small trees exhibit an exorbitant number of possible branching patterns (Harding, 1971; Van Pelt *et al.*, 1992). Given the relatively small number of observations enabled by experimental data acquisition, it is impossible to know if there is only a subset of tree shapes that fully represents a given cell type (Uemura *et al.*, 1995). It is therefore useful to have an informative scalar morphometric about tree shapes. Van Pelt *et al.* (1992) defined the *tree asymmetry index* as the average partition asymmetry across a tree, finding this measure sensitive to different tree shapes.

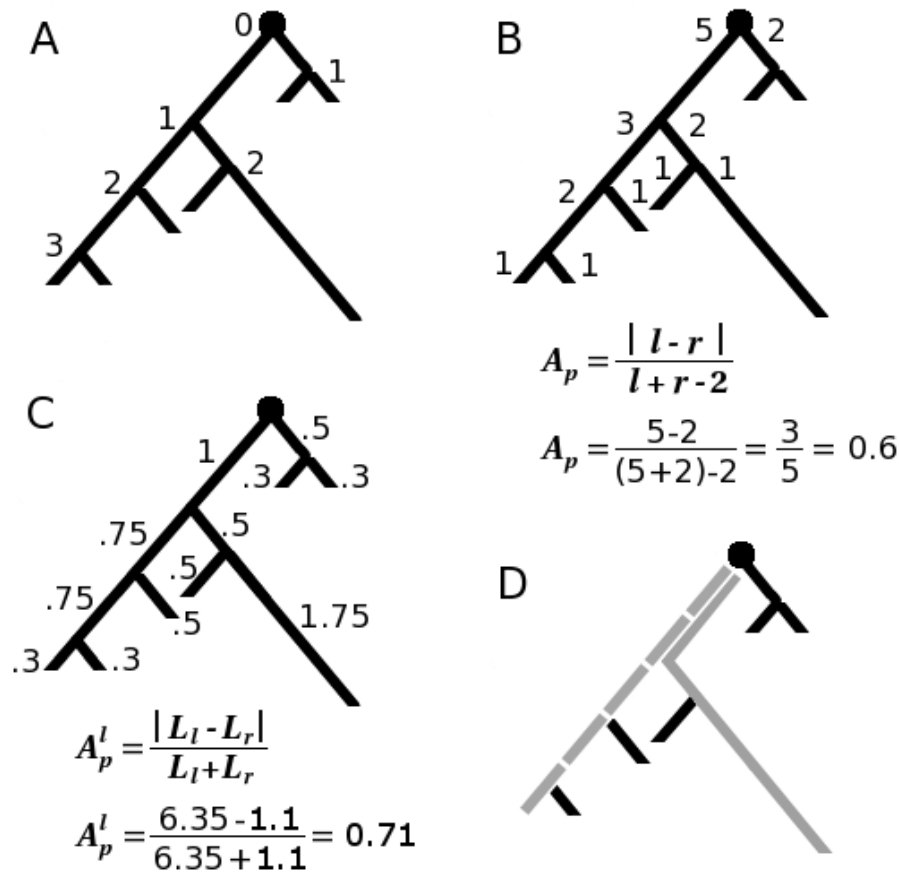


Figure A2.2. Binary tree properties. Branches are the regions between bifurcation or termination nodes. **A.** Branch (centrifugal) order of each node from the root. **B.** Terminal degree (number of terminations) of each subtree at each node. The formula for partition asymmetry (A_p) is provided below with calculation for the root of the example tree. **C.** Lengths of individual segments with length-based partition asymmetry (A_p^l) for the root. **D.** Main paths (gray) by degree (dashed) and by length (solid). At the split of the different main paths, the left subtree has a degree of 3 and a length extent of 2.6, while the right subtree has a degree of 2 and a length extent of 2.75.

Asymmetry alternatives

Because partition asymmetry measures the asymmetry of a node given all possible degree distributions, nodes that can only have one possible distribution (i.e. nodes with degree 2 or 3) technically do not have an asymmetry value. Van Pelt *et al.* (1992) chose to give the two possible partitions with degrees < 4 asymmetry values of 0

and 1 for practical purposes, but recognize that these values are arbitrary and mention several alternative methods to measure asymmetry in neuronal trees. Because arbitrary values change the average partition asymmetry for a given tree, we decided to compare asymmetry measures with and without nodes of degree < 4 . Measuring partition asymmetry for all axonal and dendritic trees from NeuroMorpho.Org with total degree > 5 ($N = 4597$), we found that excluding nodes with degrees < 4 resulted in significantly greater standard deviation ($0.19 > 0.12$, $F = 0.39$, $p < 0.0001$) suggesting higher sensitivity to tree shapes. Van Pelt *et al.* (1992) performed a similar comparison for both asymmetry measures for all possible tree shapes with degree ≤ 7 as well as measuring the number of unique values and found analogous results.

Measuring asymmetry at the tree level as opposed to locally provides a different measure that depends on node size. The metric we termed “global asymmetry” accomplishes this by weighting partition asymmetry by degree. Thus, nodes that contribute to a greater part of the entire tree have proportionally greater influence on the measured asymmetry. Since this measure is global in nature, a more absolute partition asymmetry is appropriate. This is achieved by not subtracting 2 in the denominator of the partition asymmetry equation, such that trees with a greater absolute difference in subtree sizes always have greater partition asymmetries.

Asymmetry can also be measured using continuous size metrics such as length, surface area, and volume. Since the branches diverging from a bifurcation have continuous size, even a node with two terminals can be asymmetric. There is therefore no reason for limiting the nodes to be included in such a measure. Additionally, since it is

length (or surface area or volume) and not degree being measured, degrees of freedom are not a factor (**Figure A2.2C**). Donohue and Ascoli (Donohue and Ascoli, 2008) tested the accuracy of growth models in computer simulations using different asymmetry metrics. Basal and apical tree degree asymmetries were accurately produced by different morphometric constraints (diameter and path distance, respectively). Surface area asymmetry however was best determined by path distance for both arbor types. These results suggest that different asymmetry measures may be determined by distinct factors depending on neuron or arbor type.

A novel implementation of partition asymmetry termed “excess asymmetry” was used to measure morphological homeostasis in hippocampal dendritic trees (Samsonovich and Ascoli, 2006). This metric is based on the idea that if an arbor grows under homeostatic constraints, then larger subtrees will offset smaller subtrees, limiting the size range of the entire tree. Specifically, consider a node with subtrees l and r , where l has subtrees a and b , and r has subtrees c and d . Switching b with either c or d will result in a more symmetric node in case of homeostasis (if a and b are large then c and d are small or vice versa). Excess partition asymmetry then compares the actual partition asymmetry with the swapped partition asymmetries, producing a positive value if there is homeostasis and a value near 0 otherwise. Degree, length, and surface area were all used to determine excess asymmetry. Significantly positive excess asymmetry values were found in several dendritic types of cortical principal cells, suggesting that subtrees grow under homeostatic constraints.

Asymmetry's complex relationships

Size metrics have been found to classify arbor and cell types better than some simple asymmetry scalars (Cannon *et al.*, 1999). However, asymmetry's relevance may require more targeted methods to fully capture cell type variation. Cuntz *et al.* (2007) modeled a dendritic structure by minimizing total wiring to a sample of potential synaptic sites from a reconstructed fly tangential cell. The resulting virtual dendrite was very different from the corresponding reconstructed cell. It was quite asymmetric and had a clear main path. A second model constrained both total wiring and path distance from synaptic site to soma. Diameter and taper rate were also constrained to minimize the current transfer variability between synapse to soma paths. The result appeared much like the source cell, with several levels of fairly symmetric branching followed by more asymmetric regions. A single asymmetry measure would not be capable of describing the multiple dendritic sub-domains, whereas a set of targeted measures could capture such functionally related patterns.

Neurons abide by many constraints and functions beyond homogeneous synaptic efficacy and minimizing metabolic cost. A more generalized recent model to relate neuronal function and morphology involved an evolutionary algorithm, with “genes” made up of tree growth parameters, and electrophysiological simulations to test virtual neuron fitness (Stiefel and Sejnowski, 2007). Based on the chosen functional parameters, the model finds optimal dendritic arbors over many generations. Within the resulting set of morphological parameters, those vital to the specified neuronal fitness are more highly conserved.

Asymmetry may correlate with other measurements that more directly reflect the function of the corresponding tree shape feature. The next section introduces “caulescence”, a novel measure that typically correlates with asymmetry, yet often provides a clearer functional consequence. Additionally, we explore analyses suggesting that for certain cell types, larger caulescence appears to cause the high asymmetry values.

Caulescence

The term “caulescent” was originally used in botany to describe a plant with a main trunk or stem (Martyn, 1793). As some neurons present tree structures with prominent main paths, including axons and pyramidal cell apical dendrites, we found it appropriate to develop a rule for univocally identifying main paths and a measure to capture their prominence.

The main path of a neuronal tree can be defined as the path from the soma to a termination which at each bifurcation leads to the greatest extent of a given metric (e.g. degree, length, surface area or volume). Degree-based caulescence (C_D) follows the path that leads to the higher degree at each bifurcation, while length-based caulescence (C_L) takes the path towards the most length (**Figure A2.2d**). Caulescence is then defined as the weighted partition asymmetry of nodes along the main path. This choice factors out the influence of secondary subtrees and weights more heavily the bifurcations with the largest extent. Both the partition asymmetry and weighting are based on the same metric used to determine the main path itself. The equation for caulescence is:

$$C = \sum |l - r| / \sum (l + r)$$

Equation A2.1

where l and r are the sizes of the two subtrees at each node on the main path. The resulting value represents the overall balance of extent, where high caulescence (close to 1) indicates a very distinct main path and low values reflect a more balanced tree without a clear main path. **Figure A2.3** shows main paths in example neuron types.

The significance of caulescence in neuronal arborizations is revealed in the drastic differences between the values of various cell and arborization types. Using the length metric and NeuroMorpho.Org arbors, pyramidal cell apical dendrites tend to be highly caulescent (0.62 ± 0.17 , $N=889$), while their basal counterparts have much less caulescence (0.37 ± 0.17 ; $N=1779$; $p < 0.001$). As seen in **Figure A2.3a**, the main path is more obvious in apical compared to basal dendritic trees. High caulescence by volume or surface area in dendrites can optimize synaptic integration along the main path, analogous to elongating the soma. In fact, surface area (as well as volume, not shown) caulescence is significantly higher than length caulescence for both pyramidal basal (0.39 ± 0.17 ; $p < 0.05$) and apical dendrites (0.69 ± 0.16 ; $p < 0.001$). This fits with secondary (e.g. oblique) branches having smaller diameter than main branches and therefore high input impedance (Goldstein and Rall, 1974; Vetter *et al.*, 2001; Migliore *et al.*, 2005).

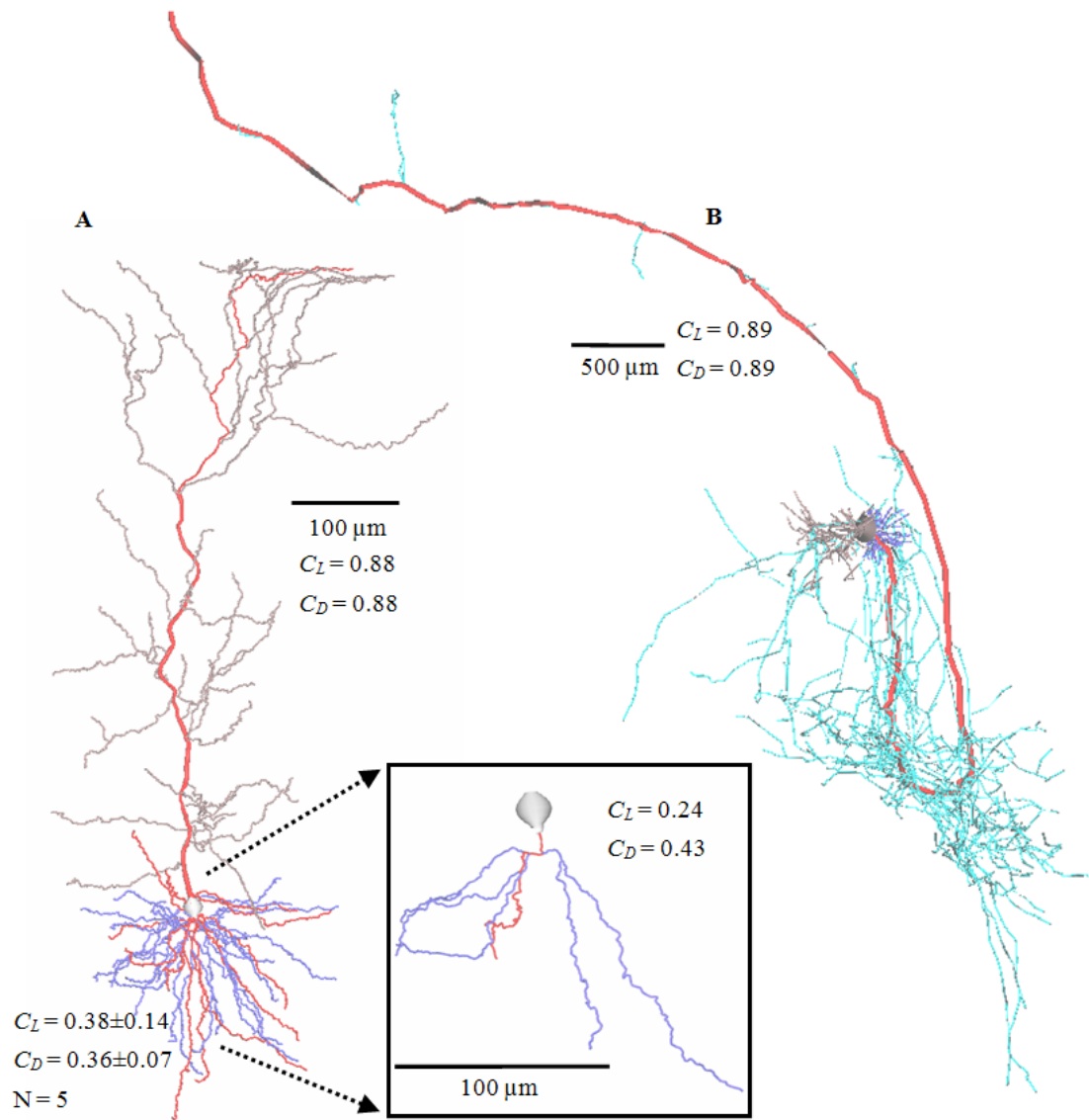


Figure A2.3. Main paths in example neurons from NeuroMorpho.Org. **A.** Hippocampal subicular pyramidal dendritic arbor (Golding *et al.*, 2001, 2005). Main path (red) is more caulescent in apical (gray, $C_L = 0.88$; $C_D = 0.88$) than basal trees (blue, $C_L = 0.38 \pm 0.14$; $C_D = 0.36 \pm 0.07$; N = 5). **Boxed Inset:** Zoomed-in (2x) individual basal tree with main path (red) ($C_L = 0.24$; $C_D = 0.43$). **B.** Hippocampal CA1 pyramidal cell dendritic (basal: dark blue; apical: gray) and axonal (light blue) arbors (Tamamaki and Nojyo, 1990, 1991; Scorcioni and Ascoli, 2005). Axonal main path (red) would not be obvious without color, even though this axonal tree is highly caulescent ($C_L = 0.89$; $C_D = 0.89$). Branches are thickened for visual clarity. C_L = length-based and C_D = degree-based caulescence.

Two groups of NeuroMorpho.Org axons had sample sizes large enough for analysis and varied substantially between types. Cortical basket cell axons had a lower mean caulescence (0.54 ± 0.15 ; $N=57$) than uniglomerular projection neuron axons (0.73 ± 0.10 ; $N=233$; $p < 0.001$). As would be expected due to the lack of taper in axons, neither cell class had axons with a significantly different surface area (C_S) or volume-based (C_V) caulescence compared to C_L and C_D . The main path in both apical dendrites and axons is acting in part as an efficient way to get from an origin (i.e. soma) to a destination (i.e. apical tuft, axonal terminals), as opposed to sending many branches to neighboring targets.

The adaptation of partition asymmetry in the caulescence definition results in a fairly strong correlation between caulescence and several asymmetry measures, particularly global asymmetry. There are, however, noticeable distinctions. For example, in apical dendrites of rodent hippocampal pyramidal cells, caulescence values below 0.50 are generally lower than, and correlate less with, global asymmetry values (both using the length metric, **Figure A2.4**). Caulescence values above 0.50 highly correlate with and fall above global asymmetry. Thus, the ratio of global asymmetry to caulescence is higher at low than at high values (**Figure A2.4** top left inset). This may be due to hippocampal apical dendrites with a major bifurcation near the soma, effectively leading to a second main path. A tree with more than one true main path has lower caulescence due to a highly weighted bifurcation with low partition asymmetry where the second main path begins. The tree's global asymmetry would then be higher than its caulescence because other highly weighted bifurcations on the second main path would contribute a relatively

high partition asymmetry. Rodent cortical apical dendrites show no significant difference in correlation above or below caulescence of 0.5 (**Figure A2.4**). While rodent hippocampal (0.66 ± 0.21 ; N=205) and cortical (0.67 ± 0.13 ; N=258) apical dendrites have similar mean caulescence, primate cortical apical dendrite values (0.57 ± 0.15 ; N=387) are substantially lower. These results signify important differences between both brain region and species even within a cell and arbor type.

Lower global asymmetry may be due to secondary branches. Measuring partition asymmetry separately along the main and secondary paths (**Figure A2.4** bottom right inset) can distinguish the relationship of potentially separate domains. A 2-sample Kolmogorov-Smirnov test rejected that the primary and secondary branch partition asymmetry values could come from the same population ($p < 0.001$). These results show that main path nodes indeed have significantly higher partition asymmetry than secondary nodes. This means that the asymmetry of apical dendrites is due primarily to the effect of the main path.

For simplicity's sake, only caulescence defined along a single main path was analyzed. Future incarnations of the metric may be expanded by taking into account the multiple main paths that certain individual trees exhibit, such as some of the low caulescence but high global asymmetry apical dendrites in **Figure A2.4**. Additionally, cerebellar climbing fibers and Purkinje cells (**Figure A2.1b**) appear to have main paths that split in multiple directions. Another aspect of separating morphological sub-domains is the determination of where a main path ends. Removing acaulescent regions would further differentiate sub-domains. For example, the apical tuft of apical dendrites is more

symmetric and thus its exclusion from the main path would likely further increase the high caulescence values of these trees.

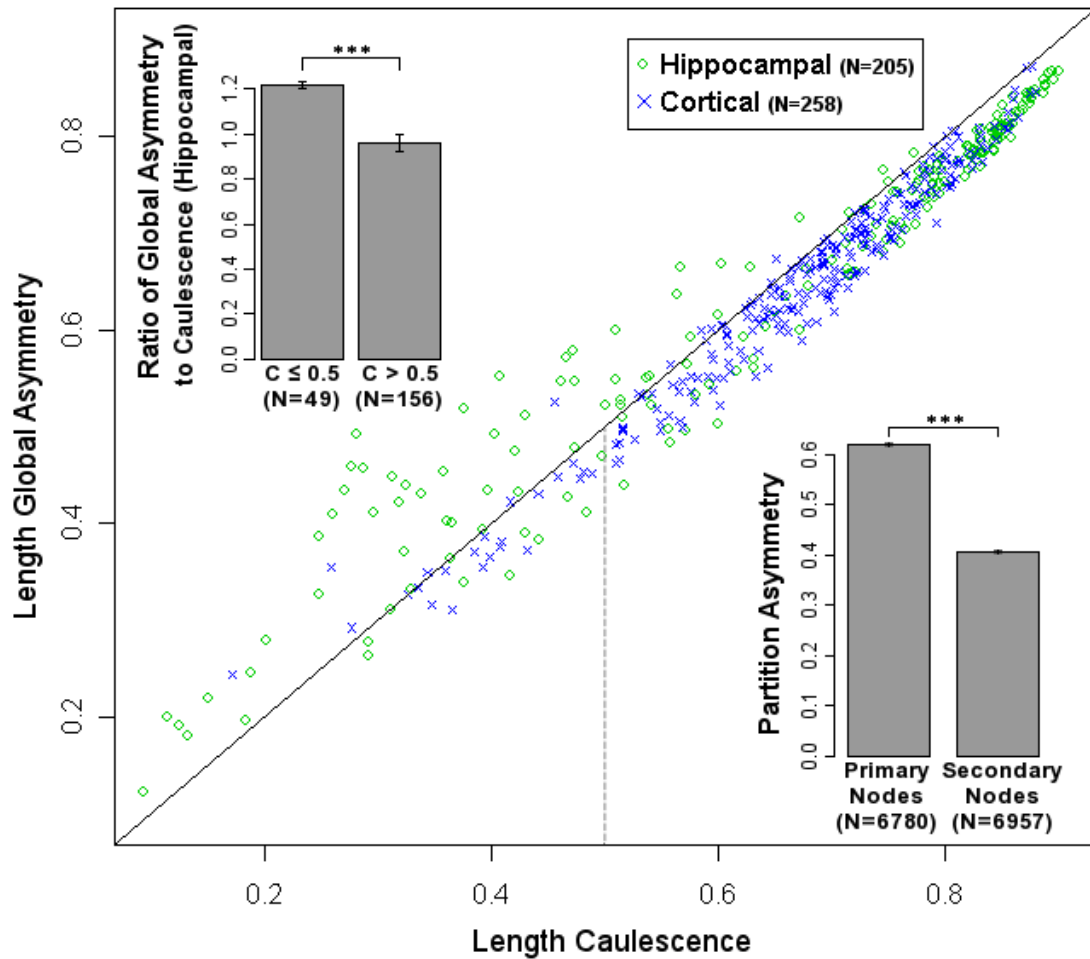


Figure A2.4. Global asymmetry versus caulescence. Scatter plot of rodent pyramidal cell apical dendrite global asymmetry against caulescence (C), both using the length metric. Green circles represent hippocampal neurons, blue crosses cortical neurons. Solid is the diagonal, where global asymmetry equals caulescence. At $C > 0.5$, caulescence is greater than global asymmetry, suggesting a difference between bifurcations in the primary and secondary paths. Variability in global asymmetry for hippocampal apical dendrites increases at $C < 0.5$. This could be explained by dendrites with multiple main paths. **Top Left Inset:** Mean ratios of global asymmetry to caulescence below and above 0.5 caulescence is significantly different for hippocampal cells ($p < 0.0001$). **Bottom Right Inset:** Partition asymmetry analysis of all rodent apical dendrite nodes. Primary path bifurcations have significantly higher partition asymmetry than secondary path bifurcations ($p < 0.0001$).

Conclusion

While many biophysical mechanisms determine cell types throughout the body, such as chemical messengers and metabolic constraints, neuronal characterization would be incomplete without considering morphology. It is both cause and effect of function; a dynamic player in signal propagation, integration, and connectivity. Size morphometrics summed through the trees can characterize fundamental aspects of different cell types. Differential size distributions at branch points serve to quantify less obvious morphological aspects such as asymmetry. All these properties are mediated by function, environment, and metabolic efficiency. Growth and functional models, in conjunction with morphological properties, have the capacity to isolate individual features and increase the understanding of their relationships to one another. One such feature, caulescence, provides a novel tool for characterizing a relatively unexplored aspect of neuromorphology.

Acknowledgements

We thank Dr. John L. Baker, Dr. Duncan E. Donohue, and Ms. Susan Wright for their constructive feedback. This work was supported by NIH grant R01 NS39600 from NINDS, NIMH, and NSF under the Human Brain Project.

APPENDIX THREE: ON COMPARING NEURONAL MORPHOLOGIES WITH THE CONSTRAINED TREE-EDIT-DISTANCE

[With kind permission from Springer Science+Business Media: *Neuroinformatics*, volume 7, 2009, Todd A. Gillette, John J. Grefenstette, © Springer Science+Business Media, LLC 2011]

Abstract

The constrained tree-edit-distance provides a computationally practical method for comparing morphologies directly without first extracting distributions of other metrics. The application of the constrained tree-edit-distance to hippocampal dendrites by Heumann and Wittum is reviewed and considered in the context of other applications and potential future uses. The method has been used on neuromuscular projection axons for comparisons of topology as well as on trees for comparing plant architectures with particular parameter sets that may inform future efforts in comparing dendritic morphologies. While clearly practical on a small scale, testing and extrapolation of run-times raise questions as to the practicality of the constrained tree-edit-distance for large-scale data mining projects. However, other more efficient algorithms may make use of it as a gold standard for direct morphological comparison.

Body

Morphological analysis of neurons has a rich history of many measurement techniques for a variety of morphological properties. Often these properties are themselves the subject of study, such as surface area and its effects on neuronal excitability (Krichmar *et al.*, 2002) or the impact of diameter and bifurcations on passive and active signal propagation (Vetter *et al.*, 2001). At other times interest is focused on the relationship between neuronal classes using a selection of morphometrics (Cannon *et al.*, 1999). Some metrics, such as Sholl analysis (Sholl, 1953) and tree asymmetry (Van Pelt *et al.*, 1992), can be used to produce distance values between neurons based on topology (i.e. the connectivity of a tree, disregarding geometrical properties such as length and diameter), but none of them produce an explicit distance value for the topologies of two neurons. A metric capable of producing such a value would be useful in determining the topological variation within and the differences between neuronal classes, particularly if individual geometric data could be factored into the differences at the branch level.

In “The Tree-edit-distance, a Measure for Quantifying Neuronal Morphology”, Heumann and Wittum’s use of the constrained tree-edit-distance between unordered trees on neuronal dendrites provides a new perspective on comparing neuronal morphologies (Heumann and Wittum, 2009). The tree-edit-distance offers a more explicit measure of similarity between two neuronal trees that actually represents the topological and geometric changes that would need to be applied to one tree in order to create the second tree. This could perhaps be considered a developmental distance, akin to the evolutionary distance derived from genetic sequence alignments. The constrained tree-edit-distance

makes the metric computationally practical. The authors illustrate the effectiveness of the method by automatically clustering sets of neurons into groups that reflect cell type. In so doing, they present a method and an application that can be used to differentiate cell classes and to measure topological differences in a more integrated manner than previous methods.

The constrained tree-edit-distance between unordered trees is a polynomial time algorithm (i.e. run time grows by some polynomial of the problem's size) developed by Kaizhong Zhang (Zhang, 1996) to improve on the original tree-edit-distance algorithm which is NP-complete (i.e. likely unsolvable in polynomial time). The tree-edit-distance is the weighted sum of the costs of a series of operations, including insertions, deletions, and geometric alterations of nodes (i.e. bifurcations and terminations), in order to make one tree the same as a second tree. This process requires that nodes from each tree be mapped to each other. The constrained version of the method imposes certain restrictions on how nodes can be mapped, described further in the article, which are responsible for the decreased computational complexity (and run-time) of the algorithm. The costs of the edit operations and the geometric measures represented at the nodes are parameters of the algorithm. They may be chosen to test a particular hypothesis or to optimize the ability of the method to differentiate groups of trees.

Heumann and Wittum investigated whether the tree-edit-distance can capture neuronal morphology and the differences between cell types. They compared neurons of four cell types: CA1 pyramidal, CA3 pyramidal, dentate gyrus granule, and dentate gyrus interneurons, with an array of label schemes. Each label scheme used a different size

attribute including length, volume, surface area, bifurcation angle, and no attribute at all (i.e. purely topological, no edit cost). Upon calculating edit distances between all pairs of neurons under a given label scheme, the authors proceeded to run a cluster analysis on each cell type pair. The battery of tests using the various schemes showed the capability of each to separate the various cell types from each other. A few label schemes quite successfully separated pyramidal cells from non-pyramidal cells, but were only partially effective at separating CA1 from CA3 and rather poor at separating granule cells from interneurons. Yet several schemes were highly effective in tests run on computer generated cell classes including several different pyramidal classes. As a whole, these results suggest that the unlabeled constrained tree-edit-distance is a useful tool for comparing neuronal morphologies.

This is not the first time that the constrained tree-edit-distance has been used on neurons. In a recent publication in PLoS Biology, a modified version of the constrained tree-edit-distance was applied to neuromuscular projections in the mouse (Lu *et al.*, 2009). The study compared left and right intra-animal connectomes in the interest of determining whether they were correlated and in some way genetically determined. Pairing axons from the left and right connectomes by motor-unit size, tree-edit-distance values were not significantly different from values produced through comparisons across animals or with “random” virtual axons generated based on neuromuscular projection branching statistics. Heumann and Wittum’s use of the tree-edit-distance is an interesting complement to the neuromuscular projection study in that the former aimed to

differentiate neuronal classes despite within class variance, while the latter looked for correlations within the variance in topology of a particular neuronal class.

Heumann and Wittum's effort is a strong first step in the development and description of the tree-edit-distance as a tool for morphological analysis and comparison of neurons. Certainly a comparison of more cell types would prove interesting. Perhaps more useful will be an exploration of node encodings with multiple dimensions, an investigation the authors recognized but stated was beyond the scope of their current effort. Such a future study could be informed by uses of the tree-edit-distance in other domains. In the domain of botany, the method has been used to compare plant architectures (Ferraro and Godin, 2000; Segura *et al.*, 2008). In Ferraro and Godin's implementation, each node is represented by multiple labels which are normalized by their mean and standard deviation (i.e. z-score) across the node's tree. Binary or class attributes were assigned heuristic cost tables (i.e. scoring matrices). They used the Manhattan distance to integrate the differences of the normalized label values. Also included was a plant specific attribute, fruit position, which in neurons may be analogous to synapse or channel location or density.

One problem with the constrained tree-edit-distance is that while it is far more efficient than the original tree-edit-distance, polynomial time algorithms can still take a long time with large data sets. With expanding numbers of neuronal morphologies available to be mined (Halavi et al., 2008), the speed of algorithms comparing neurons will become increasingly important. Heumann and Wittum found the edit distance of all neurons against all other neurons in a set. Running their program using the purely

topological labeling scheme on a Pentium 4 CPU (2.80GHz) within Cygwin produced run-times represented in **Figure A3.1**. As theory predicts, time grows by a polynomial of the total size, in branches (internal and terminal), of all of the neurons combined. A sample of 40 neurons, which with the data set used is approximately 104 branches per neuron on average and 4160 branches total, takes 492 seconds (8 minutes 20 seconds) to run.

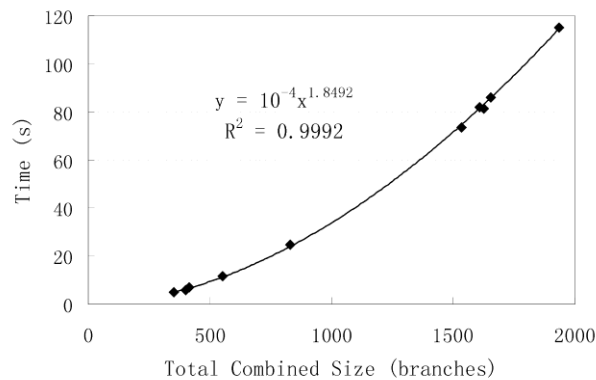


Figure A3.1. Algorithm run-time. Time taken to run the tree-edit-distance algorithm on various subsets of the Duke-Southampton archive (Cannon et al., 1999). Curve shows polynomial fit with fitted equation.

It is worthwhile to consider how this program would run on an existing open database of morphologies. The NeuroMorpho.Org database currently contains over 5000 neurons with a total of just under 320,000 branches (approximately 64 branches per neuron, though not all neurons' dendrites are traced). One may be interested in comparing a particular neuron (i.e. a query neuron) against those in a database in order to classify it or to find the most similar neurons. The time taken naturally depends on both the size of the query neuron as well as the total size of the database, as seen in **Figure A3.2**. That suggests that running a single neuron of 64 branches against that database would take about 1294 seconds (nearly 22 minutes). A better processor, particularly one

capable of running such an algorithm in parallel, would reduce this time considerably, but ultimately the practicality of the algorithm will depend on its intended use. For large scale database comparisons, it may be of interest to consider some analog of the genomic sequence comparison tools such as the Basic Local Alignment Search Tool (BLAST; Altschul et al., 1997). However, for research involving smaller sets of neurons or with enough time to let the computer process, the constrained tree-edit-distance seems a viable option.

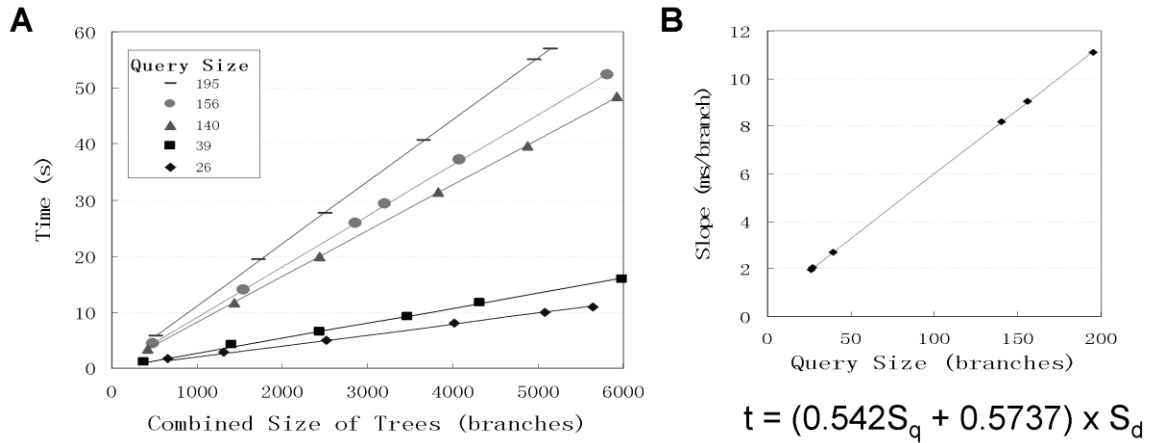


Figure A3.2. Run-time of queries by size. a. Time taken in seconds to run various query neurons of different sizes against sets of neurons from the Duke-Southampton archive. Each symbol represents one query neuron. Lines are linear fits, each with $R^2 > 0.99$. **b.** The slope (ms/branch) of fits from (a) plotted against query size. This fit produced the equation below used to estimate run-times of a query against all neurons within the NeuroMorpho.Org database. Parameters: t is time (ms), S_q is query size (branches), and S_d is the size of the database (branches).

While the constrained tree-edit-distance may or may not be immediately useful for BLAST-like searches, it can serve as a method of validation for algorithms that aim for such a purpose. As the number of morphologies increases, fast algorithms will become more important for studies that seek to mine the new wealth of data. As speed

tends to come at the cost of completeness, a gold standard will be important for determining the strengths and weaknesses of newer, faster algorithms. Moreover, there are still labeling schemes to be tested to determine the best representations. Heumann and Wittum's application of the constrained tree-edit-distance shows that it is useful for direct morphological comparison between neurons and sets the stage for future morphological investigations of neuronal classes.

Editorial Note: The authors of the target article (Heumann and Wittum, 2009) acknowledge that the commentary is correct and thank its authors. They also add that complexity issues have not been considered in the present paper, since there was no problem to run the code on the specified data. Because the algorithms can be parallelized easily, they do not see serious complexity problems in using the methods on larger datasets. Complexity issues will be the topic of a forthcoming paper.

Information Sharing Statement: The source code for the algorithm, compilation and usage instructions, and sample reconstruction files (.hoc) were provided by the authors through the Neuroinformatics editors. NeuroMorpho.Org data was retrieved through the v.3.2 database with permission and access provided by the database curator.

Acknowledgements: We would like to thank Giorgio Ascoli for his valuable discussions as well as Maryam Halavi and Deepak Ropireddy for their useful feedback.

APPENDIX FOUR: STATISTICAL ANALYSIS AND DATA MINING OF DIGITAL RECONSTRUCTIONS OF DENDRITIC MORPHOLOGIES

[Sridevi Polavaram, Todd A. Gillette, Ruchi Parekh, Giorgio A. Ascoli: in *Frontiers in Neuroanatomy* (2014) doi:10.3389/fnana.2014.00138]

Abstract

Neuronal morphology is diverse among animal species, developmental stages, brain regions, and cell types. The geometry of individual neurons also varies substantially even within the same cell class. Moreover, specific histological, imaging, and reconstruction methodologies can differentially affect morphometric measures. The quantitative characterization of neuronal arbors is necessary for in-depth understanding of the structure-function relationship in nervous systems. The large collection of community-contributed digitally reconstructed neurons available at NeuroMorpho.Org constitutes a “big data” research opportunity for neuroscience discovery beyond the approaches typically pursued in single laboratories. To illustrate these potential and related challenges, we present a database-wide statistical analysis of dendritic arbors enabling the quantification of major morphological similarities and differences across broadly adopted metadata categories. Furthermore, we adopt a complementary unsupervised approach based on clustering and dimensionality reduction to identify the main morphological parameters leading to the most statistically informative structural classification. We find that specific combinations of measures related to branching

density, overall size, tortuosity, bifurcation angles, arbor flatness, and topological asymmetry can capture anatomically and functionally relevant features of dendritic trees. The reported results only represent a small fraction of the relationships available for data exploration and hypothesis testing enabled by sharing of digital morphological reconstructions.

Introduction

The diversity of neuronal morphologies can have broad and profound functional consequences in the nervous system, which have only begun to be understood. Dendritic geometry directly impacts (and mediates) computational processes such as signal integration, coincidence detection, and logical operations (London and Häusser, 2005). The location, orientation, and shape of neural arbors enable (and strongly affect) network connectivity, providing the anatomical substrate to investigate structure-function relationship at the circuitry level (Shepherd and Svoboda, 2005; Briggman and Denk, 2006; Kajiwarra *et al.*, 2008; Weiler *et al.*, 2008; Burgalossi *et al.*, 2011; Ropireddy *et al.*, 2011; Brown *et al.*, 2012). These areas of scientific investigation apply to the morphological differences observed both within and between neuron types across animal species, developmental stages, and brain regions (**Figure A4.1**).

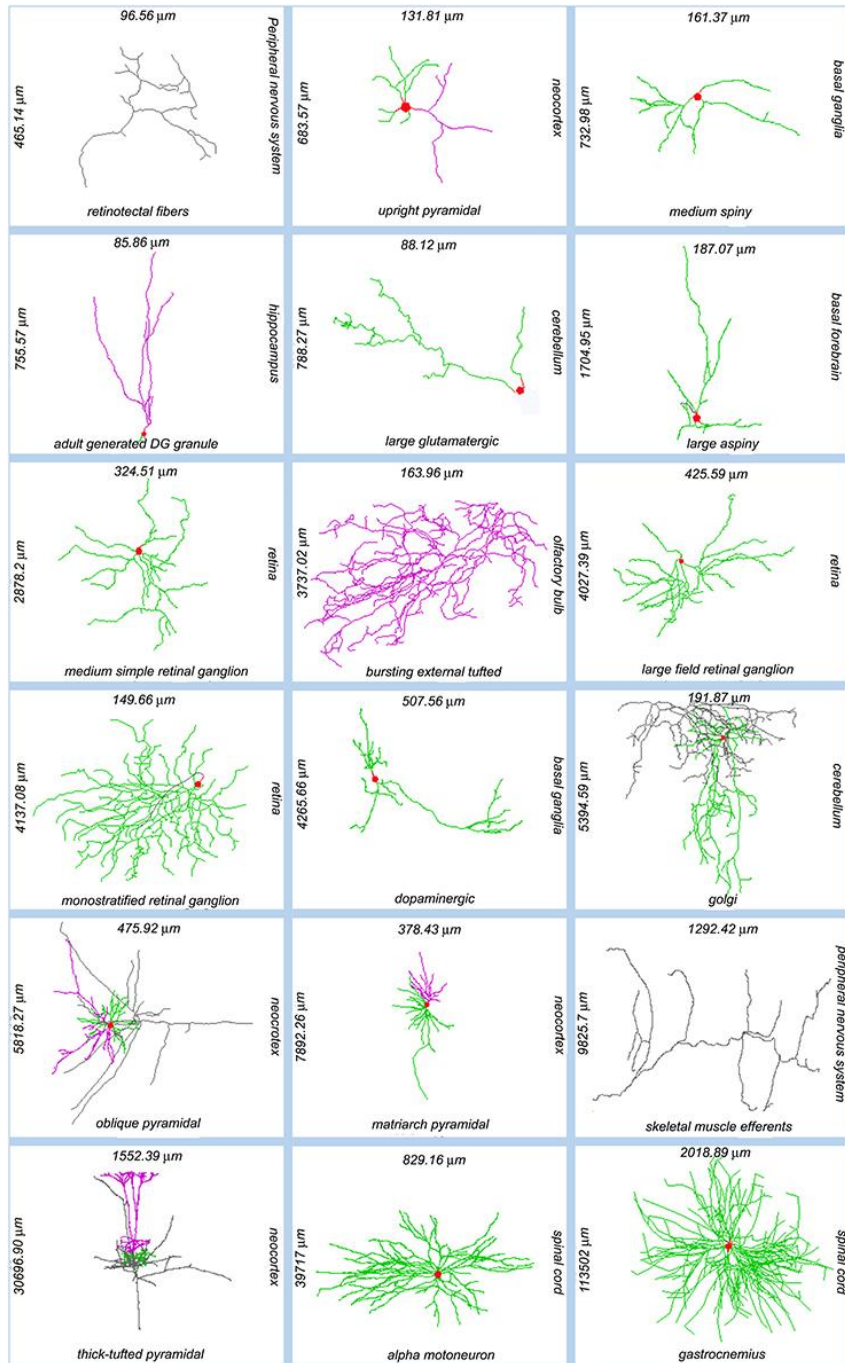


Figure A4.1. Sample of NeuroMorpho.Org reconstructions representing the anatomical diversity of dendritic and axonal trees. Each image is labeled (clockwise from its right side) with the somatic brain region, neuron types, total arbor length, and arbor width. Somata: red; axons: gray; (basal) dendrites: green; apical dendrites: magenta. NeuroMorpho.Org IDs of these neurons from left to right: 06787, 04183, 04457, 06312, 05713, 04477, 00779, 06216, 00777, 05491, 00888, 06904, 06141, 06295, 07707, 07763, 00690, 00606.

Three-dimensional digital reconstructions of axonal and dendritic arbors, combined with neuroinformatics tools and computational approaches, allow considerable opportunities for data processing, analysis, and modeling at both cellular- and systems-level (Parekh and Ascoli, 2013). The open availability of these reconstructions in databases such as NeuroMorpho.Org (**Figure A4.2**) enables re-analysis of shared data (Ascoli, 2007). As of version 5.6, the repository contained over 10,000 reconstructions contributed by 120 laboratories from 21 species, 85 brain regions and 123 cell types, representing more than 240,000 hours of manual tracing. NeuroMorpho.Org users can browse the data by animal species, brain region, cell type, and contributing lab. The “search by” option can be used to select and combine specific metadata criteria (**Figure A4.2**, left panel top) from a drop-down menu of categories such as developmental stage, experimental condition, and reconstruction method. The morphometry search functionality (**Figure A4.2**, left panel bottom) allows users to find neurons matching any combination of more than 20 morphometric criteria. From the resulting summary list of neurons (**Figure A4.2**, middle panel), individual pages for each reconstruction can be retrieved, thus displaying related metadata, a link to the associated publication, and the pre-computed morphometrics (**Figure A4.2**, right panel). Each reconstruction is downloadable as the standardized version along with the original contributed version. The log files detailing the changes made during the standardization process are available for download as well. From the individual neuron pages, users can also launch an animation file and an interactive 3D viewer.

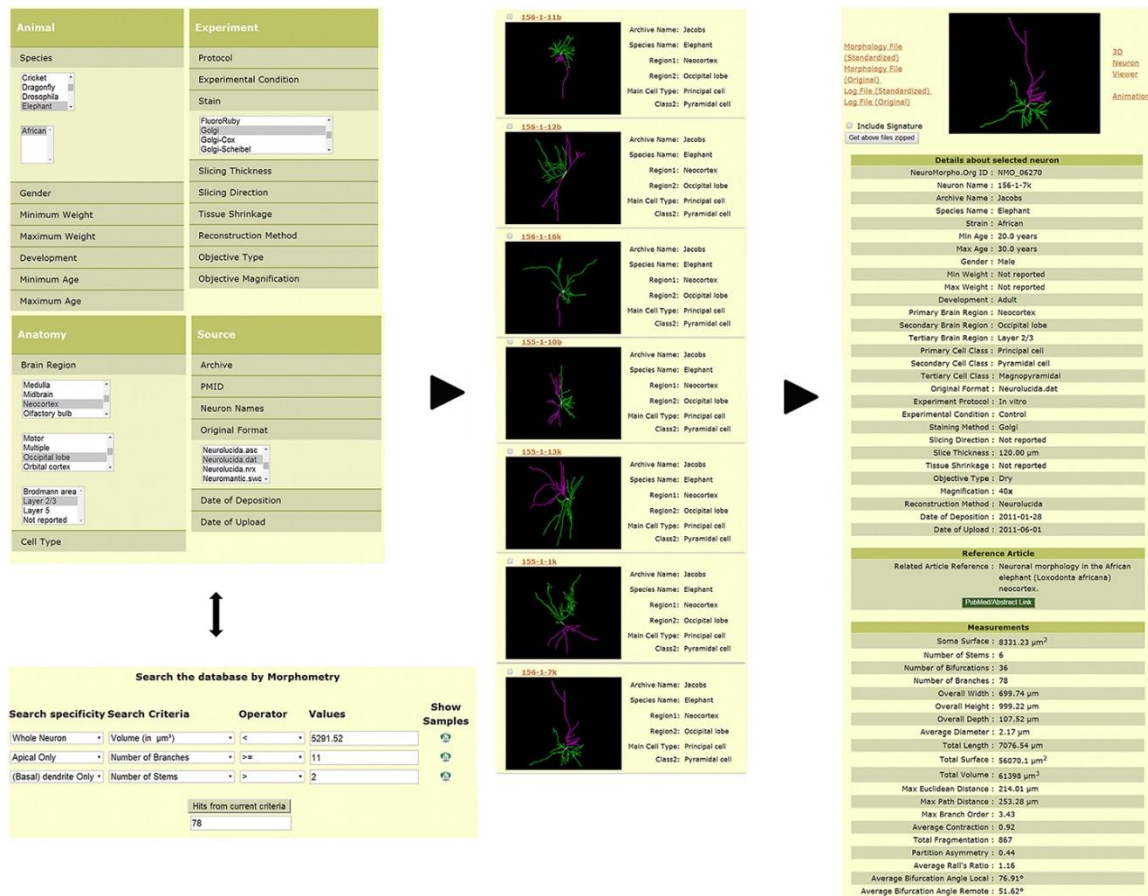


Figure A4.2: Search and download features available in NeuroMorpho.Org. Users can query the database via a number of functionalities to obtain desired reconstructions. The example provided here shows two such options. Reconstructions can be identified by selecting specific metadata across different categories such as species, brain region, cell type, staining method, and original file format (**left panel, top**). Alternatively, reconstructions can be selected by a morphometric search (**left panel, bottom**), wherein users can restrict the search to a specific arbor type (for example, apical dendrites) and define quantitative criteria to restrict particular measures (such as length or number of bifurcations) to ranges of interest. The resulting reconstructions can be displayed (among other options) with a summary of associated metadata (**middle panel**). The complete metadata and morphometric details are included within each individual neuron page (**right panel**).

Quantitative morphometry of neuronal reconstructions is often used for shape analysis (Uylings and van Pelt, 2002; Van Ooyen *et al.*, 2002; Rocchi *et al.*, 2007), also in conjunction with biologically-inspired computational simulations (Ascoli *et al.*, 2001; van Ooyen, 2011). For example, statistical distributions of morphological features are

used in stochastic growth algorithms for generating virtual trees (van Pelt *et al.*, 1997; Donohue and Ascoli, 2008; Koene *et al.*, 2009; Evans and Polavaram, 2013; Memelli *et al.*, 2013). Moreover, statistical analyses of neuronal reconstructions facilitate and support theoretical investigations. These studies for instance provided evidence for optimal wiring principles of neuronal arbors (Wen and Chklovskii, 2008) and their power law distributions, which may relate to synaptic input sampling (Lee and Stevens, 2007; Snider *et al.*, 2010; Teeter and Stevens, 2011; Cuntz *et al.*, 2012).

This study uses the L-Measure software tool (Scorcioni *et al.*, 2008) to extract morphometric data from neuronal arbors for large scale statistical analyses of available data. L-Measure computes simple statistics of morphometric features as well as their frequency distribution and inter-dependence (e.g., how arbor length varies with path distance from the soma). This tool has been used in a broad range of applications, including multidimensional analysis of neuronal shape (Zawadzki *et al.*, 2012; Costa *et al.*, 2014) and comparative studies of sensory neurons in the fly (Ting *et al.*, 2014) and of respiratory neurons in the pre-Bötzinger complex (Koizumi *et al.*, 2013). In conjunction with L-Neuron (Ascoli and Krichmar, 2000), L-Measure has also been employed to generate and validate a large-scale model of the dentate gyrus with half a million neurons (Schneider *et al.*, 2012). L-Measure has also enabled analysis of non-neuronal arbors such as arterial vasculature (Wright *et al.*, 2013), and was integrated into other digital reconstruction and analysis systems, such as the Farsight toolkit (<http://farsight-toolkit.org>) and Vaa3D (<https://code.google.com/p/vaa3d>).

With the first successes in high-throughput automatic digital neuronal tracing (Chiang *et al.*, 2011) and overall increasing volumes of published and shared reconstructions (Halavi *et al.*, 2012), “big data” opportunities for knowledge mining are starting to emerge. On the one hand, this increasing availability of shared data may foster remarkable discoveries. On the other, the heterogeneous source of data and disparate experimental conditions also pose non-trivial challenges to database-wide analyses. As a step toward large database analysis, here we utilize exploratory data analysis to quantify morphological similarities and differences across broadly diverse dendritic arbors. In the process, we recognize several critical limitations when pooling together widely non-uniform data sets. Consequently, we propose selection criteria and methodological choices aimed to maximize the potential biological relevance of the results. With such a research design, dimensionality reduction and unsupervised clustering reveal tentative morphological relationships between specific neuron types involving branching density, topology, size, and tortuosity. At the same time, we identify the most delicate factors in both data and metadata that must be considered to optimize the impact of future large-scale morphological investigations.

Methods

Selection of Datasets and Morphometric Features for Analysis

The entire pool of 10,004 reconstructions downloaded from NeuroMorpho.Org v5.6 was screened for a pre-determined set of inclusion criteria to improve interpretability of the results. Specifically, in order to be considered for analysis, digital neuron reconstructions had to (a) belong to the “control” experimental condition; (b)

contain at least four dendritic bifurcations; (c) include branch-path information and not just bifurcation connectivity; and (d) have non-zero depth range (eliminating two-dimensional tracings). The 7,143 reconstructions matching these characteristics were analyzed by their NeuroMorpho.Org metadata assignments to specific animal species, brain region, and cell type. Subsequently, for the purpose of cluster analysis chi-square testing (see below), groups of fewer than 40 neurons in any metadata combination of species, brain region, cell type, and lab of origin were excluded to ensure sufficient statistical power (Yates *et al.*, 2002). This further selection reduced the number of reconstructions to 5,099, divided into 45 unique metadata groups.

Because of the major differences between axonal and dendritic morphology, and the remarkable abundance of reconstructed dendrites relative to axons, only dendritic arbors were included in this study. Focusing on a more consistent and comparable dataset allows addressing more biologically relevant questions. Moreover, this choice reduces the errors due to incomplete reconstructions, which are considerably more severe for projection axons than for dendrites.

L-Measure allows extraction of approximately 100 distinct features from each neuron (see <http://cng.gmu.edu:8080/Lm> for full listing and detailed definitions). Of these, all measures related to branch diameter were excluded due to their strong dependence on imaging resolution, optical magnification, and other experimental details causing excessive inter-laboratory variability (Scorcioni *et al.*, 2004). All other features were subjected to cross-correlation analysis, and those with correlation greater than 80% were sequentially eliminated one at a time (re-running the cross-correlation at each step)

as they were considered highly redundant with the rest of the features. This selection left 27 features (**Table A4.1**) that were used for the remainder of the analysis. Dendritic arbor size measures consisted of total length, number of tips, height, width, and depth. Bifurcation measures included average partition asymmetry as well as amplitude, tilt, and torque angles measured locally with the adjacent tracing points or remotely with the preceding and following bifurcations or terminations. Branch measures consisted of length, tortuosity, and fractal dimension. Lastly, local measures included branch order, terminal degree, path distance from soma, and helicity.

Table A4.1. Coefficients of variation of all L-Measure derived morphometric features.

Morphometric features	CV for Dendrites	
	Hierarchy groups	Cluster groups
I. WHOLE TREE/NEURON SIZE		
Summed total arbor length	1.38	0.57
Number of arbor tips	1.65	1.82
Total arbor width	0.68	0.43
Total arbor height	0.65	0.51
Total arbor depth	1.12	0.65
II. BIFURCATION MEASURES		
Avg. partition asymmetry	0.27	0.26
Avg. local amplitude angle	0.17	0.17
Max. local amplitude angle	0.19	0.18
Avg. remote amplitude angle	0.21	0.18
Max. remote amplitude angle	0.24	0.23
Avg. local tilt angle	0.14	0.13
Max. local tilt angle	0.08	0.08
Avg. remote tilt angle	0.09	0.08
Max. remote tilt angle	0.05	0.05
Avg. local torque angle	0.17	0.16
Max. local torque angle	0.11	0.11
Avg. remote torque angle	0.18	0.17
Max. remote torque angle	0.10	0.10
III. BRANCH MEASURES		
Avg. tortuosity	0.08	0.07
Avg. fractal dimension	0.03	0.02
Max. fractal dimension	0.15	0.14
Avg. branch path length	0.59	0.41
Max. branch path length	0.81	0.53
IV. COMPARTMENT MEASURES		
Max. branch order	0.85	0.85
Avg. terminal degree	0.71	0.68
Max. path distance from soma	0.76	0.57
Max. branch helicity	0.19	0.16

A detailed description of each metric is provided at <http://cng.gmu.edu:8080/Lm/help/index.htm>.

Principal Component Analysis (PCA) and Cluster Analysis

In order to reduce the dimensionality of the morphometric space for unsupervised clustering, PCA was run on the feature dataset using the “prcomp” routine in R (v. 2.15.1). This transformation rotates all extracted measures (27 features for 5,099 arbors)

such that the first dimensions in the new space capture the most variance (in decreasing order). Prior to PCA, all features were normalized by their respective standard deviations, and the features with absolute skewness greater than unity (17/27) were log-transformed. Negatively skewed distributions were inverted and distributions with negative values were shifted prior to log-transformation. These steps ensure an approximately normal distribution of the input features, as assumed by PCA and subsequent clustering. The resulting first 17 components, accounting for 95% of the variance, were retained for cluster analysis.

Next, the dendritic arbors were clustered based on their principal morphometric components to seek a shape-based classification independent of a priori metadata grouping. We selected a model-based approach, in which mixtures of Gaussians are found that together have maximal likelihood to fit the data. A cluster is the collection of arbors that are most likely to come from the same multivariate Gaussian (referred to as a cluster model). We used the R “MCLUST” package (Fraley and Raftery, 2006) for estimating optimal model parameters and selecting the most likely model type given the dataset. The model types include spherical, ellipsoidal (with a diagonal covariance matrix), and ellipsoidal with orientation (indicating correlation between dimensions). This flexibility makes model-based clustering a more suitable choice than other popular methods (e.g., K-means) for analysis of heterogeneous data sets collated from different experiments, labs, and conditions. Not only are clusters not limited to fit spherically symmetric distributions, but also each cluster is allowed to have its own distinct variance, shape, and orientation.

MCLUST implements Expectation Maximization (EM) to select models using the Bayesian information criterion (BIC). The BIC computes the log likelihood of the cluster model, but includes a penalty for the number of parameters weighted by the log of the dataset size. Thus, goodness of fit is balanced against model simplicity according to the following equation, whereby the largest value determines the best model:

$$BIC = -2 \cdot \ln \hat{L} + k \cdot \ln(n)$$

Equation A4.1

Here, \hat{L} is the maximized likelihood computed on the marginal likelihood $P(y|M_i)$ of the candidate model M_i given the observed data y ($y_1, \dots y_n$); k is the number of free parameters to be estimated; and n is the number of data points.

The specification of MCLUST model types and parameters is coded by three letters in each of three positions. The three positions represent the model size, shape, and orientation variables, respectively. Letter “E” indicates that the parameters are equivalent across all clusters, “V” signifies variable parameter values, and “I” denotes that the corresponding parameter is not applicable. For example, “EII” indicates spherical Gaussians (no shape or orientation) with equal size among clusters, which corresponds to the traditional K-means method. Similarly, the “VVV” model type indicates varying size, shape, and orientation parameters. This latter model was determined by EM to be optimal for the data analyzed here despite its greater BIC cost implied by the larger number of free parameters. Thus, EM provides information theory-derived evidence that the performance of simple uniform spherical (K-means-like) clustering is sub-optimal for the data used in this study.

Cluster distances from the center of coordinates were measured by Z score to account for relative variance. Pairwise cluster distances were computed as the distances between the corresponding centers normalized by the cluster scatters, which are defined as averaged distance of the cluster points from the respective cluster center (Dunn, 1973). The associations among clusters and metadata groups were assessed using the chi-square test of independence, using the (marginal) frequencies of group and cluster occurrences to calculate the expected association matrix, and computing the Bonferroni-corrected p -values of the observed co-occurrences from the standardized residuals.

Results

Variability of Dendritic Morphology and Comparison by Metadata

To quantify the heterogeneity of the data, we computed the coefficient of variation (CV) for each of the 27 measured features over the entire set of 7,143 neurons as well as over the subset of 5,099 neurons used in cluster analysis (**Table A4.1**). Tortuosity, fractal dimension, and tilt angle are the least variable features, with a CV of less than 10%. In contrast, size measures are the most variable, with a CV close to or greater than unity. This apparent distinction between “local” (branch-level) vs. “global” (neuron-level) features may reflect both the effect of biological constraints (e.g., varying dimensions of different species from insects to human) and experimental conditions (slice vs. whole-animal preparations). Most other metrics display intermediate CV values.

Dendritic morphologies were then compared across species, cell types, and brain regions. The corresponding metadata information for each reconstruction in NeuroMorpho.Org was organized hierarchically (**Figure A4.3**), forming groups with a

sufficient number of neurons to enable statistical comparison of the results (at least 55 for species, 300 for brain regions, and 100 for cell types). Groups with fewer reconstructions were combined into “others” together with the reconstructions missing the detailed metadata information at the corresponding level of the hierarchy (marked as “not reported” in NeuroMorpho.Org).

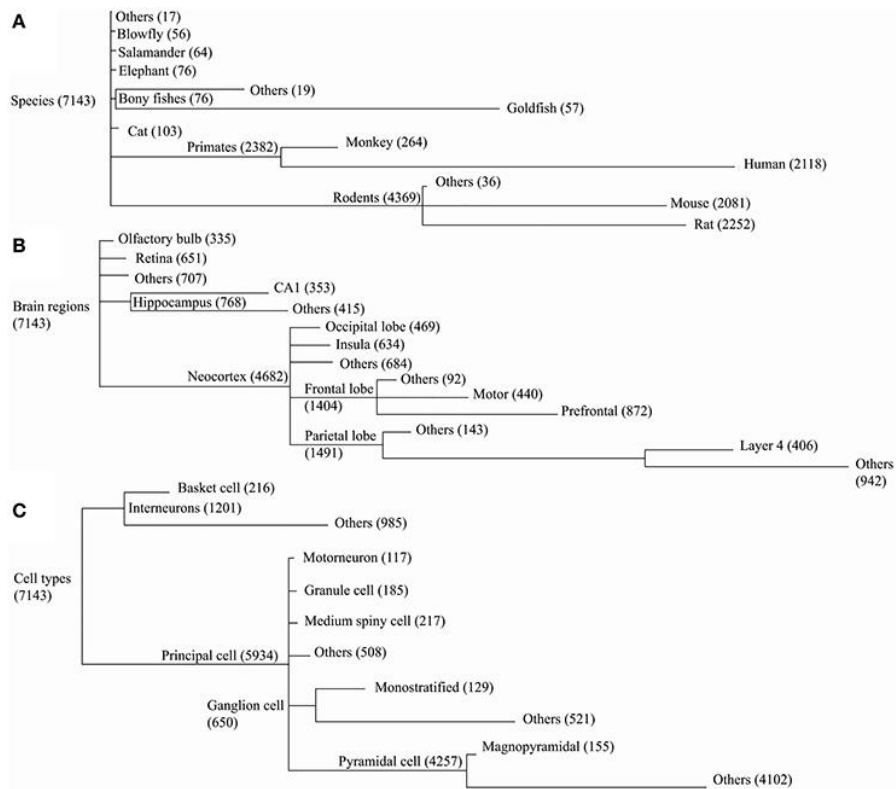


Figure A4.3. NeuroMorpho.Org v5.6 data is categorized along three major metadata dimensions, namely species, brain regions, and cell types. Reconstructions are hierarchically organized in each of these dimensions. Every node in the hierarchy is labeled by the number of associated reconstructions. The line lengths are proportional to the size of the child nodes relative to their parent nodes. **(A)** In the species hierarchy nodes with fewer than 55 reconstructions are grouped together with the “not reported” data under “Others.” In the brain regions **(B)** and cell type **(C)** hierarchies the grouping thresholds are 300 and 100, respectively.

The “leaf” nodes in each of the three metadata hierarchies (12 for species, 14 for brain regions, and 10 for cell types) were compared with a selection of representative morphometric features (**Figure A4.4**). In a limited set of cases, individual groups could be distinguished from the rest or from each other. For example, blowfly and cat reconstructions stood out against the neurons of all other species for their large topological asymmetry and Z span, respectively. The dendritic arbors of magnopyramidal cells tended to have extensive total length but low fractal dimension, whereas granule cells displayed opposite characteristics. At the same time, most groups show extensive overlap of their morphometric variance, preventing firm statistical conclusions. Part of the reason for such broad distributions is likely due to the non-uniform nature of archive-wide data sets pooled together across different experiments and laboratories. It is also clear that these metadata dimensions are not mutually independent because of evolutionary constraints (e.g., bony fishes lack a neocortex) and the finite sample of reconstructions (e.g., all monostratified ganglion cells came from the mouse retina). More generally, while popular in comparative anatomy, such a pairwise approach lacks the ability to reveal multivariate effects that are unavoidable given the non-random association between metadata groups and experimental conditions.

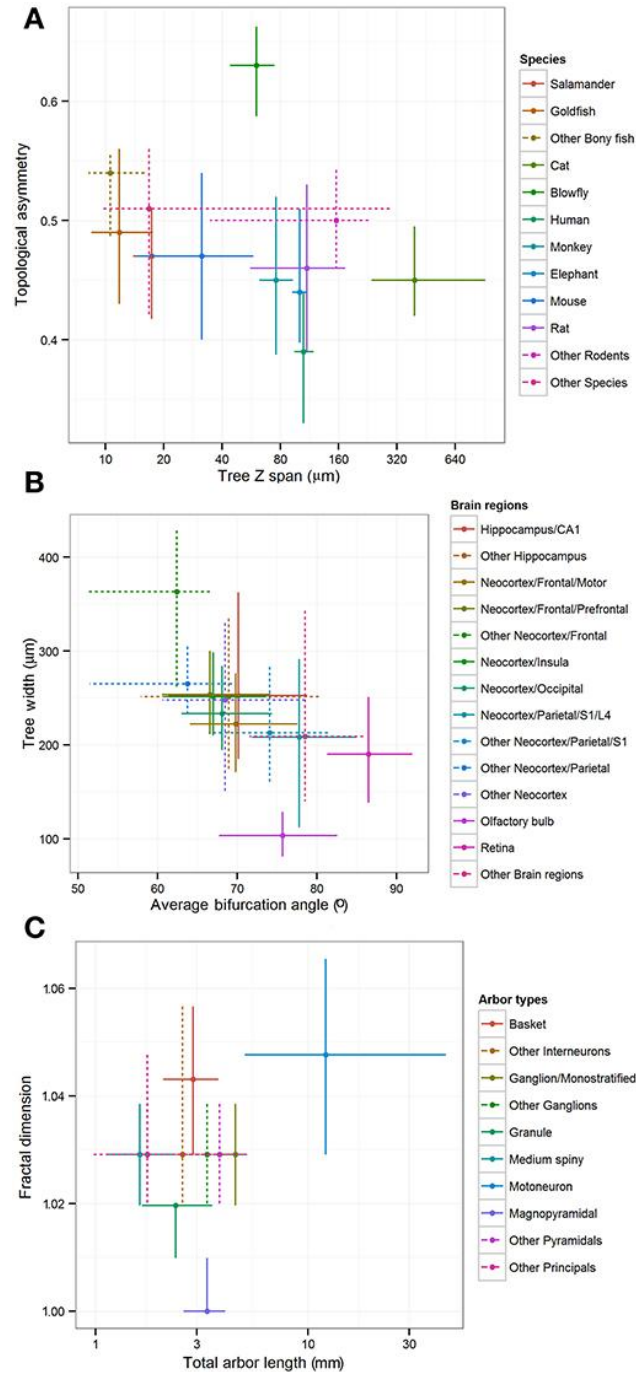


Figure A4.4. Inter-group differences of representative morphometric features within each main metadata dimension. Crosshairs represent medians and quartile ranges of each group corresponding to the leaf nodes in the hierarchies shown in Figure A4.3. Dotted lines indicate “other” groups with merged data. **(A)** Differences in arbor depth and topological asymmetry among species. **(B)** Differences in arbor width and average bifurcation angle among brain regions. **(C)** Differences in fractal dimension and total arbor length among cell types.

Extracting Primary Morphological Features by PCA and Cluster Models

In order to overcome the above limitations, we adopted an unsupervised clustering approach following dimensionality reduction with PCA. The first step is to reduce the initial parameter space to fewer orthogonal dimensions capturing most of the data variability. In mathematical terms, PCA identifies the linearly independent combinations of variables ordered by the amount of variance they explain. From the (27) original morphometric features, the first 17 dimensions of PCA covered 95% of the data variance and were used for cluster analysis.

The first 6 of these principal components were responsible for three quarters of the variance and displayed distinctive compositions of their primary morphometric features (

Table A4.2). Identifying the heaviest contributors in the linear combination of morphometric features of each principal component (“loadings”) is useful to aid subsequent interpretation of the results. The first component (PC1) is positively loaded on bifurcation angles and negatively on branch path length, reflecting high branching density. The morphometric features most descriptive of PC2 and PC3 are respectively overall size and branch tortuosity. Together, the first three components capture the majority of the data variance. The simplest morphological descriptors of PC4, PC5, and PC6 are arbor flatness (related to torque angle), fractal dimension (or “space filling”), and topological asymmetry (the average normalized sub-tree partition at bifurcation points), respectively.

Table A4.2. Morphometric loadings. Primary morphometric loading (with absolute values of 0.25 or higher) of the first six principal components of the dendritic arbors used in cluster analysis.

Principal Component	Morphometric features	Loading
PC1 (27% of cumulative variance): branching density	Max. remote amplitude angle	0.29
	Avg. remote amplitude angle	0.27
	Max. local amplitude angle	0.26
	Avg. terminal degree	0.25
	Max. branch order	0.25
	Avg. branch path length	−0.28
	Avg. remote tilt angle	−0.26
PC2 (43% of cumulative variance): size	Summed total arbor length	0.4
	Total arbor height	0.36
	Max. path distance from soma	0.34
	Total arbor width	0.33
PC3 (58% of cumulative variance): branch tortuosity	Avg. tortuosity	0.42
	Avg. fractal dimension	0.34
	Avg. local tilt angle	−0.34
PC4 (64% of cumulative variance): arbor flatness	Avg. remote torque angle	0.63
	Avg. local torque angle	0.62
PC5 (70% of cumulative variance): fractal dimension and tilt angles	Max. fractal dimension	0.37
	Avg. fractal dimension	0.35
	Avg. remote tilt angle	0.35
	Avg. tortuosity	0.25
	Max. remote tilt angle	−0.32
	Avg. remote amplitude angle	−0.36
PC6 (75% of cumulative variance): partition asymmetry and depth	Avg. partition asymmetry	0.41
	Total arbor depth	0.35

In order to produce the most informative statistical model, unsupervised clustering selects the optimal number of clusters as well as their parameters, by maximizing the BIC. These data were best fit to six clusters with varying size, shape, and orientation (**Figure A4.5**). The numerical difference between this model and the variant with constant cluster shape, however, was minimal (and is undetectable in **Figure**

A4.5a). The same model type, moreover, performed nearly as well with five or seven clusters as indicated by the absence of a clear peak in the BIC plot. We experimented with these alternative model variant and numbers of clusters and found no substantial differences in findings. At the same time, the data were *not* adequately described by traditional spherical clusters, even if with unequal sizes (**Figure A4.5a**).

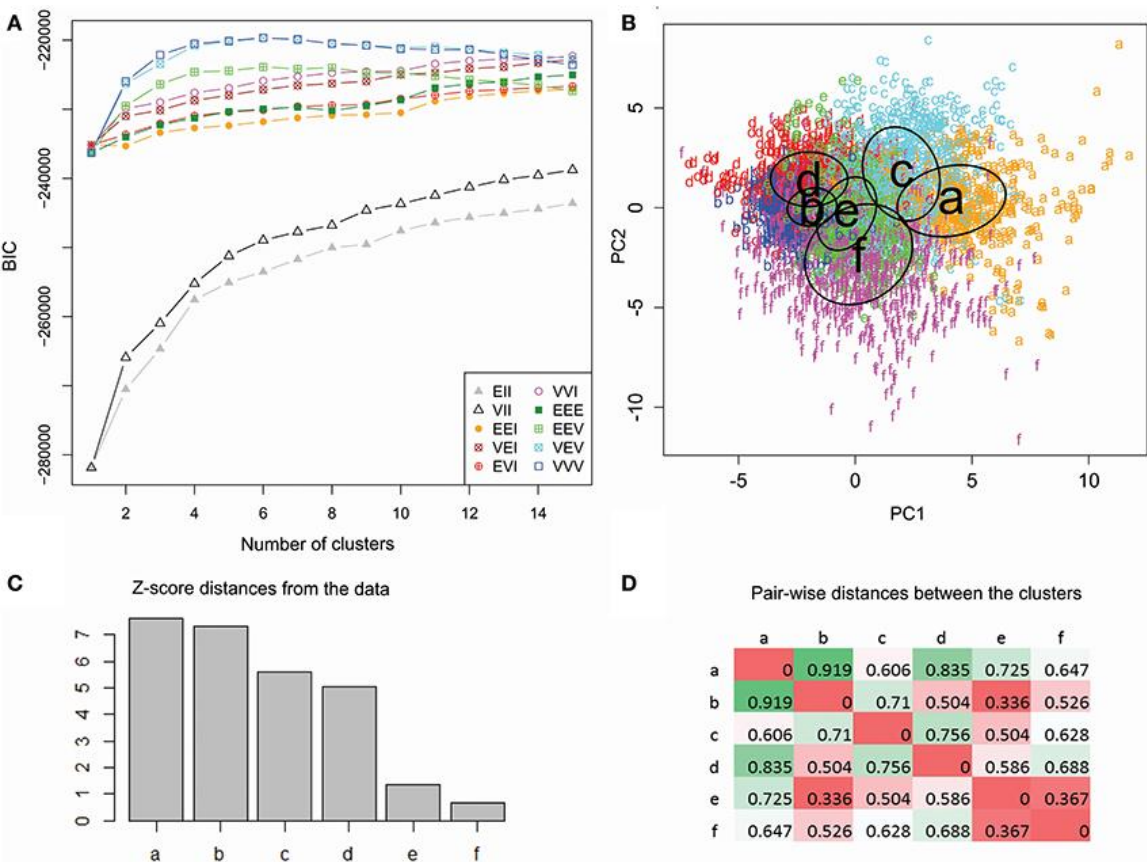


Figure A4.5. Unsupervised cluster analysis of dendritic morphology. (A) Maximization of BIC reveals marginal performance of spherical clustering with equal (EII) or unequal size (VII) alike relative to the models allowing ellipsoidal clusters. Among those, those accounting for unequal orientation (EEV, VEV, and VVV) performed better, especially with unequal size (VEV and VVV). The highest BIC value was attained at 6 clusters with varying size, shapes and orientation (VVV). (B) Scatter plot of color-coded cluster assignments (*a* through *f*) projected on the first two principal components. The ovals represent best fitting cluster parameters. (C) Cluster ranking by Z score distance from the origin of coordinates. (D) Pairwise inter-cluster distances normalized by the corresponding scatters. Farthest distances are in green and nearest are in red.

Since six clusters correspond to the maximum value for both top model types, we selected this number as the most suitable for exploratory analysis. Such a choice, nevertheless, should not be taken to reflect a ground truth that only six “true” classes exist within the data. This selection simply maximizes the inter-similarity of co-clustered classes relative to classes in other clusters given the scope, size, quality, and composition of the available dataset. To determine if further differences exist between classes that associate with the same cluster, it would be appropriate to run the same analysis on a subset of the data (sub-clustering). This additional analysis, however, requires larger datasets to meet the selection criteria based on a minimum number of reconstructions in each dataset.

The two-dimensional data projection on the first and second components illustrates the relative discrimination of clusters by branching density and arbor size (**Figure A4.5b**). Cluster ranking by variance-normalized distance from the center of coordinates (**Figure A4.5c**) allows for focused analysis on clusters farther from the origin (*a–d*), and thus morphologically distinctive, relative to those closer (*e* and *f*) to the origin. The six clusters contain respectively 585 (*a*), 1488 (*b*), 762 (*c*), 555 (*d*), 818 (*e*), and 891

(f) reconstructions. Pairwise distances (**Figure A4.5D**) reveal that one and the same cluster (*b*) is both the farthest from (*a*) and closest to (*e*) than to other clusters.

Statistical Associations between Clusters and Metadata Combinations

Unsupervised cluster models segregate neuronal reconstructions solely based on morphological features. This classification is thus complementary to, and independent of, the metadata associated with each reconstruction. The correspondence between the six morphological clusters and the 45 unique metadata groups characterized by species, brain region, neuron type, and lab of origin can shed light on the most important morphometric signatures of each metadata group. The 45-by-6 chi-square contingency matrix (**Table A4.3**) reports the probabilities that the observed over-representation and under-representations of associations between morphological clusters and metadata groups would be due to chance if the observed numerical compositions of each cluster and group were independent of each other. For example (first data row in **Table A4.3**), pyramidal neurons from mouse primary somatosensory cortex in Smit–Rigter's archive are significantly over-represented in cluster *a* ($p < 0.0002 = 10^{-3.73}$) and significantly under-represented in cluster *b* ($p < 0.001 = 10^{-3.05}$). In contrast, the proportion of these same neurons in cluster *d* is within the range expected from the sizes of this metadata group and morphological cluster.

Table A4.3. Cluster metadata associations. Matrix of positive (green) and negative (red) associations between metadata groups (rows) and morphological clusters (columns).

Metadata group (species, type, lab)	a	b	c	d	e	f
Mouse S1 pyramidal (Smit-Rigter)	3.73	-3.05	NS	25.31	NS	NS
Rat retinal ganglion (Rodger)	58.05	-4.47	NS	NS	NS	NS
Blowfly visual lobe tangential (Borst)	93.95	-3.52	NS	NS	NS	NS
Mouse retinal ganglion (Chalupa)	212.99	-15.11	NS	-3.31	-6.15	-1.52
Mouse retinal ganglion (Masland)	304.69	-23.3	-9.06	-5.8	-9.99	-9.6
Human S1 pyramidal (Jacobs)	-1.84	31.11	-2.6	NS	NS	-2.57
Human parietal lobe pyramidal (Jacobs)	NS	32.41	-2.21	NS	NS	-2.49
Human temporal lobe pyramidal (Jacobs)	NS	39.85	-1.76	NS	NS	-2.59
Human M1 pyramidal (Jacobs)	-4.94	61.98	-7.3	NS	NS	-6.02
Human V1 pyramidal (Jacobs)	-6.93	81.1	-9.97	-4.45	NS	-9.14
Human prefrontal pyramidal (Jacobs)	-14.12	196.33	-19.1	-7.79	-4.08	-18.98
Rat prefrontal pyramidal (De Koninck)	NS	-5.52	11.26	NS	NS	NS
Rat S1 pyramidal (Meyer)	NS	-5.33	25.2	NS	NS	NS
Rat frontal lobe pyramidal (Kawaguchi)	NS	-2.19	30.36	NS	NS	NS
Rat S1 pyramidal (Staiger)	NS	-2.57	32.87	NS	NS	NS
Rat S1 pyramidal (Markram)	NS	-5.74	38.27	NS	NS	NS
Mouse neocortex pyramidal (Yuste)	NS	-5.52	47.63	NS	NS	NS
Mouse S1 pyramidal (Krieger)	NS	-4.76	82.58	NS	NS	-1.58
Mouse V1 pyramidal (Yuste)	NS	-5.42	85.76	NS	-1.57	NS
Mouse S1 pyramidal (Yuste)	-3.75	-15.5	98.82	-1.96	-4.27	NS
Monkey frontal lobe pyramidal (Luebke)	NS	NS	NS	8.28	NS	NS
Rat DG granule (Claiborne)	NS	-1.66	NS	41.01	NS	NS
Monkey temp. sulcus pyramidal (Wearne_Hof)	NS	NS	NS	64.5	NS	NS
Elephant neocortex pyramidal (Jacobs)	NS	-2.23	NS	67.98	NS	NS
Monkey prefrontal pyramidal (Lewis)	-2.57	-13.46	-4.27	169.52	-4.83	NS
Human inferior frontal gyrus pyramidal (Lewis)	-2.84	-9.65	-4.47	253.7	-5.01	-5.18
Rat S1 interneuron (Helmstaeder)	NS	-1.85	NS	NS	3.96	NS
Human ant. long insular gyrus pyr. (Jacobs)	-4.34	18.18	-6.49	NS	6.57	-7.07
Human middle short insul. gyrus pyr. (Jacobs)	-4.28	18.1	-6.41	NS	8.52	-6.97
Rat M1 basket (Kawaguchi)	NS	-3.52	NS	NS	11.51	NS
Human post. short insular gyrus pyr. (Jacobs)	-4.31	11.27	-6.45	NS	13.61	-7.02
Rat S1 pyramidal (Svoboda)	NS	NS	NS	NS	17.32	NS
Rat S1 basket (Markram)	NS	NS	NS	NS	18.41	NS
Rat brainstem motoneuron (Cameron)	NS	-2.28	NS	NS	35.45	NS
Mouse M1 pyramidal (DeFelipe)	NS	-4.94	NS	NS	49.32	NS
Mouse basal ganglia med. spiny (Kellendonk)	NS	-6.66	NS	NS	75.01	NS
Mouse S1 basket (Yuste)	NS	-2.38	7.9	NS	NS	4.61
Fish retinal ganglion (Stevens)	17.87	-3.62	NS	NS	NS	6.3
Rat CA3 interneuron (Jaffe)	NS	-3.09	NS	NS	1.58	11.2
Mouse S1 interneuron (Yuste)	-1.95	-11.53	24.29	-2.22	NS	15.12
Salamander retinal ganglion (Miller)	NS	-4.28	NS	NS	NS	34.77
Rat basal forebrain large aspiny (Smith)	NS	-5.9	NS	NS	NS	64.41
Rat basal forebrain medium spiny (Smith)	-1.46	-8.47	NS	NS	NS	81.6
Mouse S1 pyramidal (Brumberg)	-1.54	-10.28	NS	-1.83	NS	97.16
Rat olfactory bulb pyramidal (Brunjes)	-4.47	-17.11	NS	-4.11	-3.92	125.81

The Bonferroni adjusted p -values obtained by the chi-square test of independence are converted for ease of comparison into \log_{10} values, inverting the sign for overrepresented (green) cells. The color gradient shows the interaction strength. Non-significant ($p > 0.05$) associations are indicated with NS.

Interestingly, each and every metadata group is over-represented in, and thus associated with, one of the six morphological clusters. The majority (38/45) are

associated with exactly one cluster, and all of the remaining (7/45) are each split between just two clusters. Most possible metadata/cluster pairs deviated significantly from the random distribution expected from the “null hypothesis”: 53 out of 270 were significantly over-represented and 87 out of 270 significantly under-represented. This overall partition of metadata groups in distinct clusters constitutes a remarkable outcome for a fully unsupervised classification method. Certain metadata groups are over-represented in one morphological cluster and under-represented in all other clusters, such as ganglion cells from mouse retina in Masland's archive (cluster *a*) and pyramidal cells from human prefrontal cortex in Jacobs' archive (cluster *b*). Other metadata groups are over-represented in one morphological cluster, but otherwise scattered throughout all other clusters per the respective numerical abundance, such as pyramidal cells from monkey frontal lobe in Luebke's archive (cluster *d*) and motoneurons from rat brainstem in Cameron's archive (cluster *e*).

Several observations can be made that transcend individual archive identities. All rodent retinal ganglion cell groups are associated with cluster *a*, whereas fish and salamander retinal ganglion cells groups are associated with cluster *f*. The relative cluster positions in the first two principal components and the corresponding morphological loadings (**Figure A4.5b** and **Table A4.2**) suggest that the retinal ganglion cells are larger and with denser branching in rodents than in non-mammals. Neocortex pyramidal cell groups are distributed across all clusters, with preference mostly based on species (most notably, human in *b*, rodents in *c*, and monkey in *d*). All rodent non-cortical and non-pyramidal cell groups are found in cluster *f* (along with salamander and fish retinal

ganglion cells). Such metadata heterogeneity, together with this cluster's minimal distance from the morphological center (**Figure A4.5c**) suggests a putative “catch-all” role for cluster *f*, which makes it broadly representative of the whole dataset.

In several cases, the split of a metadata group into two morphological clusters reflects previously reported relations. For example, three groups of pyramidal neurons from the (anterior, middle, and posterior) human insular gyrus in the Jacobs' archive divided between clusters *b* and *e* according to structural differences related to the subject's gender (Anderson *et al.*, 2009). Similarly, mouse primary somatosensory pyramidal cells are over-represented in both clusters *a* and *d*, consistent with the reported differences between young and adult animals (Smit-Rigter *et al.*, 2012). The grouping of neurons from younger mice with retinal ganglion cells (in cluster *a*) and from the older mice with pyramidal cells of larger mammals, such as monkey, elephant, and human (in cluster *d*), could be expected since the former groups are characterized by the shortest branch path length and the latter groups by the largest. The scattered clustering of pyramidal neurons, however, does not necessarily reflect existing biological relations, but might rather result from the combination of the choice of analysis algorithms, selection of parameters, and experimental differences.

The other splits of metadata groups between two clusters (**Table A4.3**) similarly revealed differences likely due to experimental procedures, such as staining protocol or slicing direction, which were not recognized in the original reports (Anderson *et al.*, 1995; Soloway *et al.*, 2002; Goldberg *et al.*, 2003; MacLean *et al.*, 2005; Nikolenko *et al.*, 2007; Woodruff *et al.*, 2009). For example, the separate clustering of different mouse

S1 pyramidal cell datasets can be explained by the differences between intracellular biocytin injection (e.g., Yuste's archive) and bulk Golgi staining (e.g., Brumberg's archive). While the mechanisms underlying the different visualization by these techniques are not yet fully understood (Thomson and Armstrong, 2011), the histological labeling information is available as metadata in NeuroMorpho.Org, thus aiding interpretation.

A complementary way to examine the associations between morphological clusters and metadata groups is to systematically analyze the composition of each cluster in terms of its associated groups, broken down by fraction of group, fraction of cluster, and neuron count (**Table A4.4**). For example (first data row in **Table A4.4**), 33% of the mouse S1 pyramidal cells from the Smit–Rigter archive are in cluster *a*, accounting for only 3% of this cluster (17 out of 560 neurons). The sums of cluster fractions in **Table A4.4** correspond to the proportion of neurons in each cluster (e.g., 97% for cluster *a*) made up by the cluster's associated metadata groups (green entries in **Table A4.3**). The remaining portions of the clusters are composed of neurons falling outside of their associated cluster. Notably, the blowfly tangential cell group is associated with cluster *a*. Moreover, clusters *b* and *c* are exclusively associated with human pyramidal cell (in which only basal dendrites are reconstructed) and rodent neocortex cell groups respectively.

Table A4.4. Composition of the six morphological clusters in terms of their over-represented metadata groups.

Cluster	Metadata group	Fraction of group	Fraction of cluster	Counts
a	Mouse S1 pyramidal (Smit-Rigter)	0.33	0.03	17
	Fish retinal ganglion (Stevens)	0.51	0.05	29
	Rat retinal ganglion (Rodger)	0.76	0.09	50
	Mouse retinal ganglion (Chalupa)	0.85	0.26	151
	Mouse retinal ganglion (Masland)	0.99	0.44	257
	Blowfly visual lobe tangential (Borst)	1	0.1	56
	Total		0.97	560
b	Human posterior short insular gyrus pyramidal (Jacobs)	0.53	0.07	106
	Human anterior long insular gyrus pyramidal (Jacobs)	0.59	0.08	118
	Human middle short insular gyrus pyramidal (Jacobs)	0.59	0.08	117
	Human S1 pyramidal (Jacobs)	0.79	0.06	95
	Human V1 pyramidal (Jacobs)	0.8	0.15	226
	Human M1 pyramidal (Jacobs)	0.8	0.12	176
	Human parietal lobe pyramidal (Jacobs)	0.86	0.06	84
	Human prefrontal pyramidal (Jacobs)	0.88	0.29	434
	Human temporal lobe pyramidal (Jacobs)	0.91	0.06	91
	Total		0.97	1447
c	Rat prefrontal pyramidal (De Koninck)	0.43	0.05	39
	Mouse S1 interneuron (Yuste)	0.47	0.09	66
	Mouse S1 basket (Yuste)	0.5	0.03	22
	Rat S1 pyramidal (Meyer)	0.6	0.06	45
	Rat S1 pyramidal (Markram)	0.66	0.07	57
	Mouse S1 pyramidal (Yuste)	0.71	0.17	128
	Mouse neocortex pyramidal (Yuste)	0.75	0.08	58
	Rat S1 pyramidal (Staiger)	0.8	0.05	37
	Rat frontal lobe pyramidal (Kawaguchi)	0.81	0.04	34
	Mouse V1 pyramidal (Yuste)	0.96	0.1	73
	Mouse S1 pyramidal (Krieger)	0.99	0.09	68
	Total		0.83	627
d	Monkey frontal lobe pyramidal (Luebke)	0.43	0.03	18
	Monkey S1 pyramidal (Smit-Rigter)	0.59	0.05	30
	Rat DG granule (Claiborne)	0.77	0.06	33
	Monkey prefrontal pyramidal (Lewis)	0.79	0.23	126
	Elephant neocortex pyramidal (Jacobs)	0.9	0.08	44
	Monkey temporal sulcus pyramidal (Wearne_Hof)	0.93	0.07	40
	Human inferior frontal gyrus pyramidal (Lewis)	0.96	0.26	146
	Total		0.78	437
e	Human anterior long insular gyrus pyramidal (Jacobs)	0.32	0.08	63
	Human middle short insular gyrus pyramidal (Jacobs)	0.33	0.08	66
	Rat CA3 interneuron (Jaffe)	0.34	0.02	20
	Human posterior short insular gyrus pyramidal (Jacobs)	0.37	0.09	74
	Rat S1 interneuron (Helmstaeder)	0.4	0.03	23
	Rat M1 basket (Kawaguchi)	0.54	0.04	30
	Rat S1 pyramidal (Svoboda)	0.58	0.05	38
	Rat S1 basket (Markram)	0.65	0.04	33
	Mouse M1 pyramidal (DeFelipe)	0.74	0.08	67
	Mouse basal ganglia medium spiny (Kellendonk)	0.83	0.1	85
	Rat brainstem motoneuron (Cameron)	0.88	0.05	38
	Total		0.66	537
f	Mouse S1 interneuron (Yuste)	0.45	0.07	63
	Fish retinal ganglion (Stevens)	0.47	0.03	27
	Mouse S1 basket (Yuste)	0.48	0.02	21
	Rat CA3 interneuron (Jaffe)	0.55	0.04	32
	Salamander retinal ganglion (Miller)	0.78	0.06	50
	Rat olfactory bulb pyramidal (Brunjes)	0.8	0.18	164
	Mouse S1 pyramidal (Brumberg)	0.88	0.13	112
	Rat basal forebrain medium spiny (Smith)	0.88	0.11	95
	Rat basal forebrain large aspiny (Smith)	0.9	0.08	73
	Total		0.72	637

Associations between metadata groups and morphological clusters are quantified as fraction of the group, fraction of the cluster, and absolute neuron count of group/cluster intersection. Within cluster, groups are arranged in ascending order of the group fraction.

Pairwise Morphometric Comparisons of Neuron Groups Identified by Cluster Analysis

Exploratory inspection of neuronal clusters in the 6-dimensional space of principal morphometric components together with the association between clusters and metadata groups (**Table A4.3, Table A4.4**) suggested closer inspection of specific morphological features in selected pairs of neuronal groups defined by their species, brain region, and cell type. The first example pertains to rodent retinal ganglion cells (**Figure A4.6**), which are characterized by high branching density and related morphological features (e.g., wide bifurcation angles). These neurons, pooled from mice and rats in four different archives, constitute 80% of cluster *a*, the farthest away from the center (**Figure A4.5c** and **Table A4.4**). At the opposite end along the first principal components is cluster *b*, entirely made of human pyramidal basal dendrites. Visual inspection (**Figure A4.6B**) reveals the distinctive shapes of rodent ganglion cells and human basal dendrites. Statistical analysis of the two main morphological loadings of PC1 (bifurcation amplitude and branch path length) confirmed the considerable difference between these two neuron groups, even when including those found in clusters other than *a* and *b* (**Figure A4.6c**).

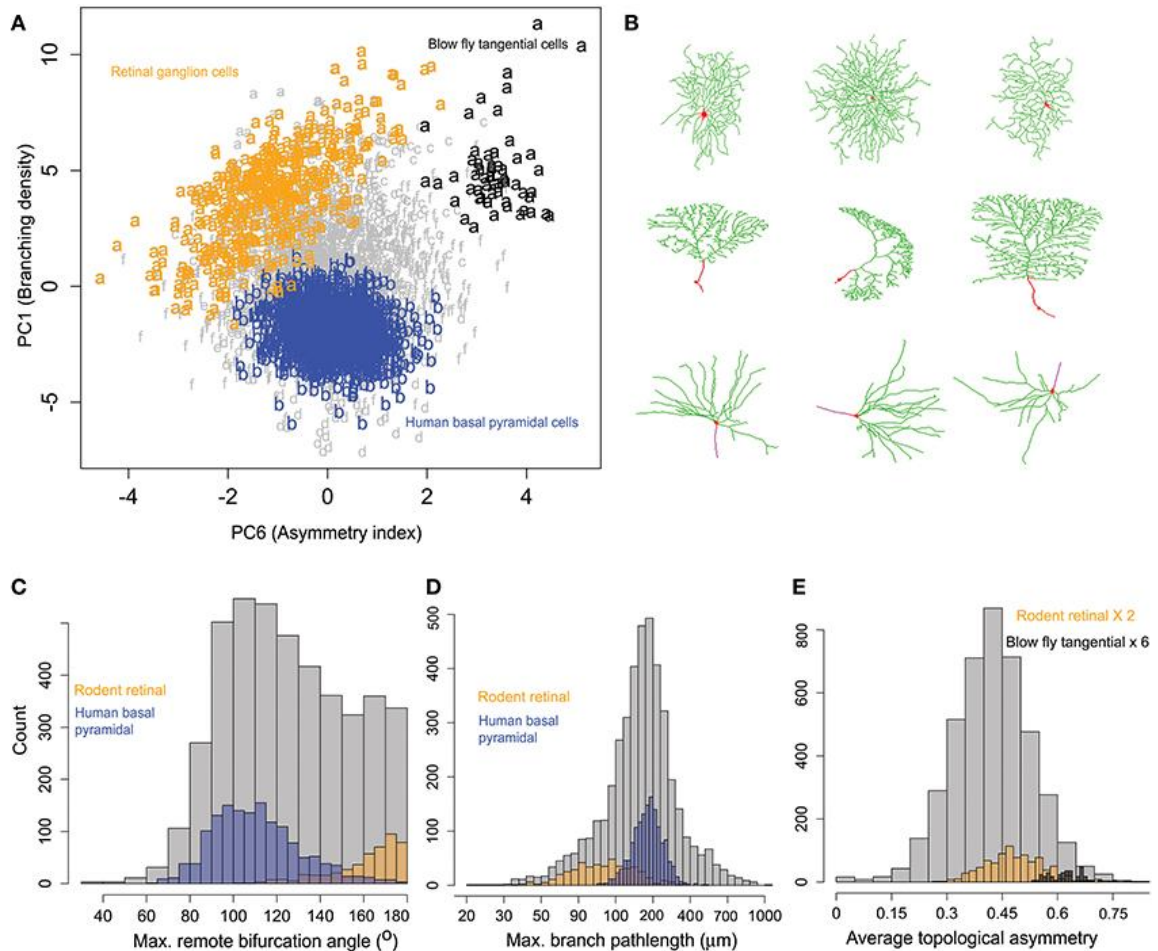


Figure A4.6. Similarities and differences of rodent retinal ganglion cells with other neurons within and across clusters. (A) All rodent retinal ganglion cells together with blowfly tangential cells (cluster *a*) show highest branching density along PC1 compared to others metadata groups. The human basal pyramidal cell cluster (*b*) is highlighted for comparison. PC6 separates the tangential and ganglion cells, showing sub-cluster differences. The retinal cells also show a pattern of increasing partition asymmetry with increasing branching density. (B) Sample images of retinal ganglion cells (top), blowfly tangential cells (middle), and human basal pyramidal cells (bottom). NeuroMorpho.Org IDs of these neurons from left to right: 06464, 05352, 05405, 06652, 01895, 06640, 03723, 03724, 03722. (C) Rodent ganglion cells have larger amplitude angles compared to human basal pyramidal cells (and most other cell classes). (D) Rodent ganglion cells also display shorter branch length, corresponding to higher branching density. (E) The blowfly neurons, while sharing similar branch path length and amplitude angles with the retinal cells, have higher topological asymmetry.

The second most prominent group in cluster *a* is constituted by blowfly tangential sensory neurons. These neurons share with the rodent ganglion cells not only comparable branching density properties captured by PC1 (low branch path length and high

bifurcation angle), but also similar distributions on PC2 through PC5 and all corresponding morphological features loading on those dimensions. These include measures of size (e.g., total dendritic length and spanned volume), of space filling (fractal dimension and tortuosity), and of arbor planarity (torque and tilt angles). Such tight alignment on the first five principal components along with the morphological co-clustering suggests a structural basis for the functional commonalities between blowfly tangential cells and retinal ganglion cells, both of which process motion-sensitive visual information (Kong *et al.*, 2005; Cuntz *et al.*, 2008).

Nevertheless, rotation on the sixth principal component exposed a surprising and nearly perfect separation between retinal ganglion cells and blowfly tangential cell (**Figure A4.6a**). Since the main morphological feature loading on PC6 is topological asymmetry (the average partition of terminal degree over all bifurcations), we compared the distribution of this measure between the two neuron classes (**Figure A4.6d**). This analysis demonstrated that blowfly tangential neurons have much more asymmetric bifurcations than ganglion cells (and most typical neurons). Interestingly, the data projection over the first and sixth principal components (**Figure A4.6a**) also suggested a linear relationship between topological asymmetry and branching density in rodent retinal ganglion cells but not in other groups. The Pearson correlation coefficients for branching density and asymmetry index ($R = -0.50$) and for bifurcation amplitude remote and asymmetry ($R = 0.51$) are both statistically highly significant ($p < 10^{-10}$).

Rotating the data along the first and third principal components (related to branching density and tortuosity, respectively) revealed another distinct relationship

across pyramidal cells from different species, brain regions, and developmental stages (**Figure A4.7**). Specifically, neocortical pyramidal cells from rodents (clusters *c*) and primates (cluster *d*) display a trend of increasing branch tortuosity with increasing branch density (**Figure A4.7a**). Visual examination of morphologies selected from the corresponding clusters in the PC1-PC3 scatter plot demonstrates a correspondence in the increase of branch density and branch tortuosity (**Figure A4.7b**). The least tortuous trees, and many of the primate neurons, are noted to be incomplete reconstructions, in which only dendrites proximal to the soma are traced. In contrast, the dendrites of rodent neocortical pyramidal neurons tend to be fully reconstructed in both apical and basal arbors.

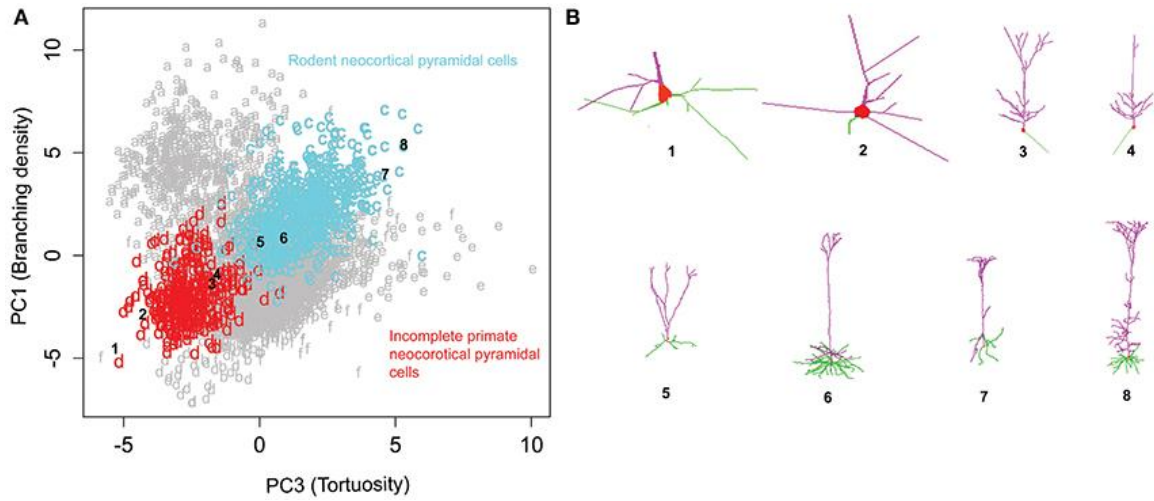


Figure A4.7. Rodent and primate cortical pyramidal cells show a distinct linear relationship between PC1 and PC3. (A) The majority (71%) of cluster c consists of rodent cortical pyramidal cells, whereas a similar proportion of cluster d (72%) corresponds to primate pyramidal cells, which tend to be only partially reconstructed. (B) Sample images of incomplete primate pyramidal cells in the top row (1–4) and rodent cortical pyramidal cells in the bottom (5–8). The numbers indicate their corresponding position in the cluster plot illustrating the progressive increase in branching density and tortuosity in both clusters. The NeuroMorpho.Org IDs of these neurons from left to right: 01821, 01526, 01627, 01623, 09630, 09474, 02569, 00266.

Critical Assessment of Potential Confounds

In the course of the iterative process of data inspection, hypothesis formulation, research design, and quantitative analysis, we encountered numerous challenges pertaining to data validation, curation, and standardization across labs. After a preliminary exploration of the entire content of NeuroMorpho.org, we decided to include in our study only approximately half of the available neurons. Specifically, we chose to avoid multi-lab analysis of axons, because of the extreme dependence of axonal morphology on experimental conditions. In our early analysis attempt that did not segregate axons from dendrites, biological findings became practically impossible to disentangle from major artifacts. This selection effectively defines a standard of minimal requirements for effectively comparing neural arbors.

Moreover, we excluded measures related to branch diameter (branch power ratios, surface areas, occupied volume, etc.) due to their strong sensitivity on the inter-laboratory variety of labeling or staining, imaging resolution or optical magnification, and other experimental details affecting tracing conditions (Scorcioni *et al.*, 2004). Furthermore, most reconstructed cells originate from preparations in acute brain slices (*in vitro*). In the primary somatosensory region of rat neocortex (S1), this common preparation may result

in trimming off more than 50% of the dendritic arbor (Oberlaender *et al.*, 2012). These slicing artifacts impact larger brains to a greater extent, as reflected by the fact that human cells are only represented by basal dendrites. In addition to species differences, trimming effects also depend on animal age, slicing thickness and orientation, and the depth of electrode penetration in the tissue. For these reasons, when mining the cluster analysis results, we paid particular attention to only report findings as “biological” (**Figure A4.6, Figure A4.7**) that were not based on size or any morphometrics significantly affected by trimming artifacts. Instead, we identified correlations based on measures such as branching density, tortuosity, and branch angles, all of which have been previously found to be consistent between *in vitro* and *in vivo* preparations (Pyapali *et al.*, 1998).

On the one hand, this judicious design allowed the independent reproduction of findings reported in prior publications. These included several cases of “split metadata groups” into two morphological clusters, which reflected structural differences related to the subject's gender (Anderson and Martin, 2009) or developmental stage (Smit-Rigter *et al.*, 2012). On the other hand, experimental artifacts still contributed to clustering, and other splits of metadata groups between two clusters (**Table A4.3**) revealed differences likely due to staining protocol or slicing direction, which were not recognized or discussed in the original reports (Anderson *et al.*, 1995; Soloway *et al.*, 2002; Goldberg *et al.*, 2003; MacLean *et al.*, 2005; Nikolenko *et al.*, 2007; Woodruff *et al.*, 2009). Thus, database-wide analyses can reveal potential confounds not easily pinpointed in individual studies.

One of the most common artifacts of tissue processing is shrinkage, and this factor is also highly variable among labs. Shrinkage differentially affects the slice planar and perpendicular dimensions (the latter typically producing a larger effect). Thinner slices tend to shrink more and so do preparations from younger animals. The duration of the experimental procedure may also impact shrinkage, as do the bathing and embedding media. Shrinkage can be measured in all dimensions and it can therefore be compensated for by multiplying the resulting position coordinates by an appropriate correction factor. However, this post-processing operation also exacerbates noise due to light diffraction and other experimental errors. These sources of errors tend to be larger in the direction corresponding to the depth of the slice (“Z”), which is usually estimated through a piezo-controller in the motorized stage. Moreover, shrinkage typically varies both within and between sections, and an accurate calibration therefore requires multiple repeated measurements that add to the already demanding labor intensity of digital reconstruction. For these reasons, shrinkage is not always measured, reported or corrected for. This variability across published studies further worsens the numerous sources of differences due to experimental processing.

In light of the above consideration, we specifically looked for potential shrinkage-related confounds in the clustering results. Out of 56 unique combinations of clusters, metadata groups, and corresponding published articles, only 14 reported shrinkage estimates or mentioned shrinkage altogether. Of those, a mere 5 applied the corresponding correction to the data. Unsurprisingly given the limited sample, we found no statistically significant association between both corrected or uncorrected values and

clustering. Next, we examined slicing thickness, which was reported in 49 (out of 56) cases (with median 200 μm). Values varied broadly from 80 to 400 μm , with 85% of them falling between 120 and 350 μm . No statistical association was found between clustering and these values. The lack of explicit shrinkage information prevents firm conclusions and leaves open the possibility that some of the findings we report may be ultimately due to slicing artifacts. However, the low coefficient of variation of measurements typically sensitive to shrinkage, especially tortuosity and fractal dimension (**Table A4.1**), suggests that the noise related to shrinkage (as opposed to that affecting size measures) may affect most of the analyzed data to a similar degree.

Fully assessing the potential usefulness of the reported results will require additional investigation. For example, morphologically detailed electrophysiological simulations might be useful to explore how the observed relations between datasets (**Figure A4.6**) or between morphological variables (**Figure A4.7**) could affect input/output relationship of individual neurons (e.g., Scorcioni et al., 2004; Komendantov and Ascoli, 2009). Similarly, the effect of these morphological relations on potential network connectivity could be studied by embedding the digital reconstructions in an appropriate three-dimensional model of the surrounding neural tissue (e.g., Chiang et al., 2011; Ropireddy and Ascoli, 2011). The continuous expansion of the available pool of neuronal reconstructions will also enable the future validation and refinement of these results with additional or independent datasets.

Discussion

This work illustrates how shared morphological data can lead to new observations of potential neurobiological interest by enabling statistical quantification of commonalities and differences among neuron groups. However, our results also demonstrate the challenges of working with large-scale datasets from heterogeneous sources, even after extensive effort in metadata curation and management as well as in data standardization and selection. Direct analysis of selected morphometric features among large neuron groups organized by the main metadata dimensions of species, brain region, and cell type failed to reveal meaningful patterns beyond the well-known variability of neuronal shape. At the same time, systematic pairwise examination of all 45 neuronal groups with distinct species, brain region, cell type, and lab of origin for each of the 27 main morphological features would produce more than 50,000 comparisons, raising questions of scientific interpretation and statistical significance.

To overcome these issues, we adopted principal component analysis to identify the most discriminant morphological features throughout the dataset, and model-based cluster analysis to segregate neuron groups solely on the basis of the morphometric characteristics. This approach allowed rigorous examination of the statistical associations between clusters and metadata and inspection of the most informative morphological measurements on the basis of their principal component loadings. The results revealed morphological differences between specific cell types and animal species that were robust to lab provenance while retaining considerable sensitivity to developmental stages and fine regional location as well as to the original experimental conditions. For example, neocortical pyramidal cells from rodents and primates alike display a trend of increasing

branch tortuosity with increasing branch density (**Figure A4.7a**). This distinct relationship, holding across different species, brain regions, and developmental stages, appears robust to slicing artifacts as demonstrated by the co-alignment of both partially reconstructed and fully reconstructed neurons (**Figure A4.7b**).

The primary features of dendritic morphology corresponded to branching density, size, space filling, and bifurcation asymmetry. Of these features, size is likely to be the most dramatically impacted by differential trimming artifacts from brains of varying size. Nevertheless, the most interesting biological findings were based on branch- or bifurcation-level observations. Rodent retinal ganglion cells stood out for their extreme branching density, and clustered together with other neuron types involved in primary sensory processing as well as with developing pyramidal cells from the somatosensory cortex of 6–9 day-old rat. Moreover, the results also highlighted species differences within the same cell types by differentiating retinal cells of rodent from those of fish and amphibians. Specifically, ganglion cells have denser branching and wider bifurcation angles in rodents than in non-mammalian vertebrates (**Figure A4.5b, Figure A4.6, Table A4.2**). This observation is based on pooling of mice and rats data from four different labs in one cluster, and of fish and salamander from two different labs in the other, and we failed to find any methodological reasons that could explain these morphological differences.

Blowfly tangential sensory neurons are similar to the rodent ganglion cells in many morphological features (e.g., low branch path length, comparable fractal dimension, tortuosity, and arbor planarity), possibly providing a geometric correlate for

their similar function in processing motion-sensitive visual information (Kong *et al.*, 2005; Cuntz *et al.*, 2008). Nevertheless, retinal ganglion cells and blowfly tangential cells can also be neatly distinguished due to the much more asymmetric bifurcations of the latter neurons (**Figure A4.6A**) relative to those of the former (and of most typical neurons). Interestingly, cluster analysis also suggested a linear relationship between topological asymmetry and branching density in rodent retinal ganglion cells but not in other groups, pointing to a previously unrecognized peculiar morphological signature of this class only.

The branching density of mature cortical pyramidal cells, in contrast, was at the opposite end relative to ganglion cells (also demonstrating the effect of developmental changes) and displayed a distinctive correlation with branch tortuosity. Adult neocortex pyramidal cells represent the largest population in NeuroMorpho.Org and come from a broad range of animals, anatomical subregions, layers, and experimental conditions, enabling certain morphological differentiations (e.g., rodent S1 vs. primate M1). Non-cortical neurons, including striatal, olfactory, and others, were distinguished for the smaller size and larger variability of their dendritic arbors.

Several recent investigations have adopted similar analysis designs and strategies for dimensionality reduction, mainly for the purpose of exploratory neuron type classification (e.g., Kong *et al.*, 2005; McGarry *et al.*, 2010; Santana *et al.*, 2013). Alternative approaches to develop automated machine-learning classifiers for identifying neuron types also promise to be effective for large data sets. The present exploratory study used multivariate morphometric analysis to identify the most informative

morphological features that distinguish between neuron groups organized by their metadata. We predict that statistical morphometric mining will also prove to be useful for developing quantitative hypotheses and for designing computational models of dendritic growth. At the same time, we discussed the considerable challenge of pooling together data from disparate experimental conditions, and the resulting analysis limitations.

Generation of standardized morphological data across laboratories and research designs could yield much more powerful large-scale data mining. In particular, we are convinced that better clustering would result from more consistent data collection. Systematic reliability assessment of experimental protocols can maximize morphological reproducibility and minimize tracing artifacts (e.g., Dercksen et al., 2014). Any such improvements would also help refine cluster analysis by reducing variability. Unfortunately, the arguably “ideal” experimental choices (*in vivo* labeling, reconstructions at the resolution limit of light, systematic measurement and compensation of tissue shrinkage, serial tracing across histological sections, etc.) also correspond to the most labor-intensive conditions for manual or semi-manual morphological reconstructions. This tension between quality, sample size, and research cost underscores the need and desirability of fully automated and broadly applicable tracing technologies (Brown *et al.*, 2011; Donohue and Ascoli, 2011).

Acknowledgments

We thank Dr. Diek Wheeler, Dr. Rubén Armañanzas, and Mr. David Hamilton for feedback on an earlier version of the manuscript. This work was supported in part by NIH grant R01 NS39600 and a Keck NAKFI award from the National Academy of Science.

Publication of this article was funded in part by the George Mason University Libraries
Open Access Publishing Fund.

REFERENCES

- Allen Institute for Brain Science (2009) Allen Brain Atlas: Home.
- Altschul SF *et al.* (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Altschul SF *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Amthor FR, Oyster CW (1995) Spatial organization of retinal information about the direction of image motion. *Proc Natl Acad Sci U S A* 92:4002–4005.
- Anderson JC, Martin KAC (2009) The synaptic connections between cortical areas V1 and V2 in macaque monkey. *J Neurosci* 29:11283–11293.
- Anderson K *et al.* (2009) The morphology of supragranular pyramidal neurons in the human insular cortex: a quantitative Golgi study. *Cereb Cortex* 19:2131–2144.
- Anderson SA *et al.* (1995) Synchronous development of pyramidal neuron dendritic spines and parvalbumin-immunoreactive chandelier neuron axon terminals in layer III of monkey prefrontal cortex. *Neuroscience* 67:7–22.
- Ascoli GA (1999) Progress and perspectives in computational neuroanatomy. *Anat Rec* 257:195–207.
- Ascoli GA *et al.* (2001) Computer generation and quantitative morphometric analysis of virtual neurons. *Anat Embryol (Berl)* 204:283–301.
- Ascoli GA (2002) Neuroanatomical algorithms for dendritic modelling. *Netw Comput Neural Syst* 13:247–260.
- Ascoli GA (2006) Mobilizing the base of neuroscience data: the case of neuronal morphologies. *Nat Rev Neurosci* 7:318–324.
- Ascoli GA (2007) Successes and Rewards in Sharing Digital Reconstructions of Neuronal Morphology. *Neuroinformatics* 5:154–160.
- Ascoli GA *et al.* (2008) Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nat Rev Neurosci* 9:557–568.
- Ascoli GA *et al.* (2009) Quantitative morphometry of electrophysiologically identified CA3b interneurons reveals robust local geometry and distinct cell classes. *J Comp Neurol* 515:677–695.
- Ascoli GA, Donohue DE, Halavi M (2007) NeuroMorpho.Org: A Central Resource for Neuronal Morphologies. *J Neurosci* 27:9247–9251.
- Ascoli GA, Krichmar JL (2000) L-Neuron: A Modeling Tool for the Efficient Generation and Parsimonious Description of Dendritic Morphology. *Neurocomputing* 32-33:1003–1011.
- Baas PW, Deitch JS, Black MM, Banker GA (1988) Polarity orientation of microtubules in hippocampal neurons: uniformity in the axon and nonuniformity in the dendrite. *Proc Natl Acad Sci U S A* 85:8335–8339.

- Baloyannis SJ (2009) Dendritic pathology in Alzheimer's disease. *J Neurol Sci* 283:153–157.
- Bartos M, Vida I, Jonas P (2007) Synaptic mechanisms of synchronized gamma oscillations in inhibitory interneuron networks. *Nat Rev Neurosci* 8:45–56.
- Bastmeyer M, O'Leary D (1996) Dynamics of target recognition by interstitial axon branching along developing cortical axons. *J Neurosci* 16:1450–1459.
- Basu S, Condrón B, Acton ST (2011) Path2Path: Hierarchical path-based analysis for neuron matching. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp 996–999. IEEE.
- Berger B, Peng J, Singh M (2013) Computational solutions for omics data. *Nat Rev Genet* 14:333–346.
- Berry M, Bradley PM (1976) The application of network analysis to the study of branching patterns of large dendritic fields. *Brain Res* 109:111–132.
- Berry M, Flinn R (1984) Vertex Analysis of Purkinje Cell Dendritic Trees in the Cerebellum of the Rat. *Proc R Soc B Biol Sci* 221:321–348.
- Bhaskar L, Krishnan VS, Thampan RV (2007) Cytoskeletal elements and intracellular transport. *J Cell Biochem* 101:1097–1108.
- Bille P (2005) A survey on tree edit distance and related problems. *Theor Comput Sci* 337:217–239.
- Binzegger T, Douglas RJ, Martin KAC (2004) A quantitative map of the circuit of cat primary visual cortex. *J Neurosci* 24:8441–8453.
- Borst A, Haag J (1996) The intrinsic electrophysiological characteristics of fly lobula plate tangential cells: I. Passive membrane properties. *J Comput Neurosci* 3:313–336.
- Bower JM, Beeman D (2003) The Book of GENESIS: second edition. New York: Springer-Verlag.
- Bremner JD *et al.* (1997) Magnetic resonance imaging-based measurement of hippocampal volume in posttraumatic stress disorder related to childhood physical and sexual abuse--a preliminary report. *Biol Psychiatry* 41:23–32.
- Briggman KL, Denk W (2006) Towards neural circuit reconstruction with volume electron microscopy techniques. *Curr Opin Neurobiol* 16:562–570.
- Brown K, Donohue D, D'Alessandro G, Ascoli G (2005) A cross-platform freeware tool for digital reconstruction of neuronal arborizations from image stacks. *Neuroinformatics* 3:343–359.
- Brown KM *et al.* (2011) The DIADEM data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions. *Neuroinformatics* 9:143–157.
- Brown KM, Gillette TA, Ascoli GA (2008) Quantifying neuronal size: summing up trees and splitting the branch difference. *Semin Cell Dev Biol* 19:485–493.
- Brown KM, Sugihara I, Shinoda Y, Ascoli GA (2012) Digital morphometry of rat cerebellar climbing fibers reveals distinct branch and bouton types. *J Neurosci* 32:14670–14684.
- Brudno M *et al.* (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics* 19:i54–i62.

- Brugg B, Matus A (1991) Phosphorylation determines the binding of microtubule-associated protein 2 (MAP2) to microtubules in living cells. *J Cell Biol* 114:735–743.
- Buchfink B, Xie C, Huson DH (2014) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.
- Burgalossi A *et al.* (2011) Microcircuits of functionally identified neurons in the rat medial entorhinal cortex. *Neuron* 70:773–786.
- Burke R, Marks W, Ulfhake B (1992) A parsimonious description of motoneuron dendritic morphology using computer simulation. *J Neurosci* 12:2403–2416.
- Cameron WE *et al.* (1991) Morphometric analysis of phrenic motoneurons in the cat during postnatal development. *J Comp Neurol* 314:763–776.
- Cannon RC, Wheal H V, Turner DA (1999) Dendrites of classes of hippocampal neurons differ in structural complexity and branching patterns. *J Comp Neurol* 413:619–633.
- Capowski J (1983) An automatic neuron reconstruction system. *J Neurosci Methods* 8:353–364.
- Cardona A *et al.* (2010) Identifying neuronal lineages of *Drosophila* by sequence analysis of axon tracts. *J Neurosci* 30:7538–7553.
- Caserta F *et al.* (1995) Determination of fractal dimension of physiologically characterized neurons in two and three dimensions. *J Neurosci Methods* 56:133–144.
- Chen J, Kanai Y, Cowan NJ, Hirokawa N (1992) Projection domains of MAP2 and tau determine spacings between microtubules in dendrites and axons. *Nature* 360:674–677.
- Chiang A-S *et al.* (2011) Three-dimensional reconstruction of brain-wide wiring networks in *Drosophila* at single-cell resolution. *Curr Biol* 21:1–11.
- Chitwood RA, Hubbard A, Jaffe DB (1999) Passive electrotonic properties of rat hippocampal CA3 interneurons. *J Physiol* 515:743–756.
- Chklovskii DB (2000) Optimal Sizes of Dendritic and Axonal Arbors in a Topographic Projection. *J Neurophysiol* 83:2113–2119.
- Chklovskii DB, Mel BW, Svoboda K (2004) Cortical rewiring and information storage. *Nature* 431:782–788.
- Chklovskii DB, Schikorski T, Stevens CF (2002) Wiring Optimization in Cortical Circuits. *Neuron* 34:341–347.
- Cline HT (2001) Dendritic arbor development and synaptogenesis. *Curr Opin Neurobiol* 11:118–126.
- Conde C, Caceres A (2009) Microtubule assembly, organization and dynamics in axons and dendrites. *Nat Rev Neurosci* 10:319–332.
- Coombs J, van der List D, Wang G-Y, Chalupa LM (2006) Morphological properties of mouse retinal ganglion cells. *Neuroscience* 140:123–136.
- Costa M *et al.* (2014) NBLAST: Rapid, sensitive comparison of neuronal structure and construction of neuron family databases. Cold Spring Harbor Labs Journals.
- Cuntz H, Borst I, Segev I (2007) Optimization principles of dendritic structure. *Theor Biol Med Model* 4:21.

- Cuntz H, Forstner F, Borst A, Häusser M (2011) The TREES toolbox--probing the basis of axonal and dendritic branching. *Neuroinformatics* 9:91–96.
- Cuntz H, Forstner F, Haag J, Borst A (2008) The Morphological Identity of Insect Dendrites. *PLoS Comput Biol* 4:e1000251.
- Cuntz H, Mathy A, Häusser M (2012) A scaling law derived from optimal dendritic wiring. *Proc Natl Acad Sci U S A* 109:11014–11018.
- Dailey ME, Smith SJ (1996) The Dynamics of Dendritic Structure in Developing Hippocampal Slices. *J Neurosci* 16:2983–2994.
- De Paola V *et al.* (2006) Cell type-specific structural plasticity of axonal branches and boutons in the adult neocortex. *Neuron* 49:861–875.
- DeFelipe J *et al.* (2013) New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nat Rev Neurosci* 14:202–216.
- Deisseroth K, Schnitzer MJ (2013) Engineering approaches to illuminating brain structure and dynamics. *Neuron* 80:568–577.
- Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B* 39:1–38.
- Dent EW, Tang F, Kalil K (2003) Axon guidance by growth cones and branches: common cytoskeletal and signaling mechanisms. *Neurosci* 9:343–353.
- Dercksen VJ, Hege H-C, Oberlaender M (2014) The Filament Editor: an interactive software environment for visualization, proof-editing and analysis of 3D neuron morphology. *Neuroinformatics* 12:325–339.
- Díez-Guerra FJ, Avila J (1993) MAP2 phosphorylation parallels dendrite arborization in hippocampal neurones in culture. *Neuroreport* 4:419–422.
- Donohue DE, Ascoli GA (2005a) Models of neuronal outgrowth. In: *Databasing the Brain: From Data to Knowledge* (Koslow SH, Subramaniam S, eds), pp 304–323. John Wiley and Sons, Inc.
- Donohue DE, Ascoli GA (2005b) Local Diameter Fully Constrains Dendritic Size in Basal but not Apical Trees of CA1 Pyramidal Neurons. *J Comput Neurosci* 19:223–238.
- Donohue DE, Ascoli GA (2008) A Comparative Computer Simulation of Dendritic Morphology. *PLoS Comput Biol* 4:e1000089.
- Donohue DE, Ascoli GA (2011) Automated reconstruction of neuronal morphology: an overview. *Brain Res Rev* 67:94–102.
- Donohue DE, Scorcioni R, Ascoli GA (2002) Generation and Description of Neuronal Morphology Using L-Neuron: A case study. In, pp 49–70. Totowa, NJ: Humana Press.
- Dotti CG, Sullivan CA, Banker GA (1988) The establishment of polarity by hippocampal neurons in culture. *J Neurosci* 8:1454–1468.
- Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J Cybern* 3:32–57.
- Eccles JC, Ito M, Szentágothai J (1967) *The Cerebellum as a Neuronal Machine*. Heidelberg, Berlin, Gottingen, New York: Springer-Verlag.
- Evans RC, Polavaram S (2013) Growing a garden of neurons. *Front Neuroinform* 7:17.

- Ferraro P, Godin C (2000) A distance measure between plant architectures. *Ann For Sci* 57:445–461.
- Fiala BA, Joyce JN, Greenough WT (1978) Environmental complexity modulates growth of granule cell dendrites in developing but not adult hippocampus of rats. *Exp Neurol* 59:372–383.
- Fischer M, Kaech S, Knutti D, Matus A (1998) Rapid actin-based plasticity in dendritic spines. *Neuron* 20:847–854.
- Fraley C, Raftery AE (2002) Model-based Clustering, Discriminant Analysis and Density Estimation. *J Am Stat Assoc* 97:611–631.
- Fraley C, Raftery AE (2006) MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering. Seattle, WA.
- Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597.
- Freund TF (2003) Interneuron Diversity series: Rhythm and mood in perisomatic inhibition. *Trends Neurosci* 26:489–495.
- Fukata Y, Kimura T, Kaibuchi K (2002) Axon specification in hippocampal neurons. *Neurosci Res* 43:305–315.
- Gao F-B (2007) Molecular and cellular mechanisms of dendritic morphogenesis. *Curr Opin Neurobiol* 17:525–532.
- Gardner D *et al.* (2008) The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience. *Neuroinformatics* 6:149–160.
- Georges PC, Hadzimichalis NM, Sweet ES, Firestein BL (2008) The Yin–Yang of Dendrite Morphology: Unity of Actin and Microtubules. *Mol Neurobiol* 38:270–284.
- Gillette T, Grefenstette J (2009) On Comparing Neuronal Morphologies with the Constrained Tree-edit-distance. *Neuroinformatics* 7:191–194.
- Gillette TA *et al.* (2011a) DIADEMchallenge.Org: A Compendium of Resources Fostering the Continuous Development of Automated Neuronal Reconstruction. *Neuroinformatics* 9:303–304.
- Gillette TA, Ascoli GA (2015) Topological characterization of neuronal arbor morphology via sequence representation. I. Motif analysis. *BMC Bioinformatics* In Press.
- Gillette TA, Brown KM, Ascoli GA (2011b) The DIADEM metric: comparing multiple reconstructions of the same neuron. *Neuroinformatics* 9:233–245.
- Gillette TA, Hosseini P, Ascoli GA (2015) Topological characterization of neuronal arbor morphology via sequence representation. II. Global alignment. *BMC Bioinformatics* In Press.
- Glaser EM, Vanderloos H (1965) A semi-automatic computer-microscope for the analysis of neuronal morphology. *IEEE Trans Biomed Eng* 12:22–31.
- Glaser JR, Glaser EM (1990) Neuron imaging with Neurolucida – a PC-based system for image combining microscopy. *Comput Med Imaging Graph* 14:307–317.
- Goldberg J *et al.* (2006) From Dendrites to Networks: Optically Probing the Living Brain Slice and Using Principal Component Analysis to Characterize Neuronal

- Morphology. In: Neuroanatomical Tract-Tracing 3: Molecules, Neurons, and Systems (Zaborszky L, Wouterlood FG, Lanciego JL, eds), pp 452–476. Springer US.
- Goldberg JH, Tamas G, Aronov D, Yuste R (2003) Calcium Microdomains in Aspiny Dendrites. *Neuron* 40:807–821.
- Golding NL *et al.* (2005) Factors mediating powerful voltage attenuation along CA1 pyramidal neuron dendrites. *J Physiol* 568:69–82.
- Golding NL, Kath WL, Spruston N (2001) Dichotomy of Action-Potential Backpropagation in CA1 Pyramidal Neuron Dendrites. *J Neurophysiol* 86:2998–3010.
- Goldstein SS, Rall W (1974) Changes of Action Potential Shape and Velocity for Changing Core Conductor Geometry. *Biophys J* 14:731–757.
- Graham B, van Ooyen A (2006) Mathematical modelling and numerical simulation of the morphological development of neurons. *BMC Neurosci* 7 Suppl 1:S9.
- Graham BP, van Ooyen A (2004) Transport limited effects in a model of dendritic branching. *J Theor Biol* 230:421–432.
- Groh A *et al.* (2010) Cell-type specific properties of pyramidal neurons in neocortex underlying a layout that is modifiable depending on the cortical area. *Cereb Cortex* 20:826–836.
- Grueber WB, Jan LY, Jan YN (2002) Tiling of the Drosophila epidermis by multidendritic sensory neurons. *Development* 129:2867–2878.
- Grueber WB, Sagasti A (2010) Self-avoidance and tiling: Mechanisms of dendrite and axon spacing. *Cold Spring Harb Perspect Biol* 2:a001750.
- Gulyás AI, Megías M, Emri Z, Freund TF (1999) Total number and ratio of excitatory and inhibitory synapses converging onto single interneurons of different types in the CA1 area of the rat hippocampus. *J Neurosci* 19:10082–10097.
- Halavi M, Hamilton KA, Parekh R, Ascoli GA (2012) Digital reconstructions of neuronal morphology: three decades of research trends. *Front Neurosci* 6:49.
- Hao H, Shreiber DI (2007) Axon Kinematics Change During Growth and Development. *J Biomech Eng* 129:511–522.
- Harding EF (1971) The Probabilities of Rooted Tree-Shapes Generated by Random Bifurcation. *Adv Appl Probab* 3:44–77.
- Härdle W, Simar L (2007) Applied multivariate statistical analysis, 2nd ed. Berlin, Heidelberg: Springer-Verlag.
- Hartenstein V (2011) Morphological diversity and development of glia in Drosophila. *Glia* 59:1237–1252.
- Hattori D, Millard SS, Wojtowicz WM, Zipursky SL (2008) Dscam-Mediated Cell Recognition Regulates Neural Circuit Formation. *Annu Rev Cell Dev Biol* 24:597–620.
- Haug H (1987) Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *Am J Anat* 180:126–142.

- He HY, Cline HT (2011) Diadem X: automated 4 dimensional analysis of morphological data. *Neuroinformatics* 9:107–112.
- Hely TA, Graham B, van Ooyen A (2001) A computational model of dendrite elongation and branching based on MAP2 phosphorylation. *J Theor Biol* 210:375–384.
- Henze DA, Cameron WE, Barrionuevo G (1996) Dendritic morphology and its effects on the amplitude and rise-time of synaptic signals in hippocampal CA3 pyramidal cells. *J Comp Neurol* 369:331–344.
- Heumann H, Wittum G (2009) The Tree-Edit-Distance, a Measure for Quantifying Neuronal Morphology. *Neuroinformatics* 7:179–190.
- Hillman D (1979) Neuronal shape parameters and substructures as a basis of neuronal form. In: pp 477–498. Cambridge: MIT Press.
- Hillman DE (1988) Parameters of dendritic shape and substructure: intrinsic and extrinsic determination? In: Intrinsic determinants of neuronal form and function (Lasek RJ, Black MM, eds), pp 83–113. New York: Liss.
- Hines ML *et al.* (2004) ModelDB: A Database to Support Computational Neuroscience. *J Comput Neurosci* 17:7–11.
- Hirokawa N (1998) Kinesin and Dynein Superfamily Proteins and the Mechanism of Organelle Transport. *Science* (80-) 279:519–526.
- Hirokawa N, Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nat Rev Neurosci* 6:201–214.
- Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544.
- Horton AC *et al.* (2005) Polarized Secretory Trafficking Directs Cargo for Asymmetric Dendrite Growth and Morphogenesis. *Neuron* 48:757–771.
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218.
- Hughes ME *et al.* (2007) Homophilic Dscam Interactions Control Complex Dendrite Morphogenesis. *Neuron* 54:417–427.
- Ikegaya Y, Le Bon-Jego M, Yuste R (2005) Large-scale imaging of cortical network activity with calcium indicators. *Neurosci Res* 52:132–138.
- Ishizuka N, Cowan WM, Amaral DG (1995) A quantitative analysis of the dendritic organization of pyramidal cells in the rat hippocampus. *J Comp Neurol* 362:17–45.
- Jaeger D (2001) Accurate reconstruction of neuronal morphology. In: Computational Neuroscience: Realistic Modeling for Experimentalists (De Schutter E, ed), pp 159–178. CRC Press.
- Jan Y-N, Jan LY (2003) The Control of Dendrite Development. *Neuron* 40:229–242.
- Jefferis GSXE *et al.* (2007) Comprehensive maps of Drosophila higher olfactory centers: spatially segregated fruit and pheromone representation. *Cell* 128:1187–1203.
- Jeffress LA (1948) A place theory of sound localization. *J Comp Physiol Psychol* 41:35–39.
- Jelinek HF, Fernandez E (1998) Neurons and fractals: how reliable and useful are calculations of fractal dimensions? *J Neurosci Methods* 81:9–18.
- Jenett A *et al.* (2012) A GAL4-driver line resource for Drosophila neurobiology. *Cell Rep* 2:991–1001.

- Kajiwarara R *et al.* (2008) Convergence of entorhinal and CA3 inputs onto pyramidal neurons and interneurons in hippocampal area CA1--an anatomical study in the rat. *Hippocampus* 18:266–280.
- Kalil K, Dent EW (2014) Branch management: mechanisms of axon branching in the developing vertebrate CNS. *Nat Rev Neurosci* 15:7–18.
- Kaspirzhny A V, Gogan P, Horcholle-Bossavit G, Tyc-Dumont S (2002) Neuronal morphology data bases: morphological noise and assesment of data quality. *Netw Comput Neural Syst* 13:357–380.
- Kasthuri N, Lichtman JW (2010) Neurocartography. *Neuropsychopharmacology* 35:342–343.
- Kawaguchi Y, Karube F, Kubota Y (2006) Dendritic branch typing and spine expression patterns in cortical nonpyramidal cells. *Cereb Cortex* 16:696–711.
- Kiddie G *et al.* (2005) Biologically plausible models of neurite outgrowth. *Prog Brain Res* 147:67–80.
- Knoblauch A, Palm G (2001) Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks* 14:763–780.
- Koene RA *et al.* (2009) NETMORPH: a framework for the stochastic generation of large scale neuronal networks with realistic neuron morphologies. *Neuroinformatics* 7:195–210.
- Koester S, O’Leary D (1992) Functional classes of cortical projection neurons develop dendritic distinctions by class-specific sculpting of an early common pattern. *J Neurosci* 12:1382–1393.
- Koizumi H *et al.* (2013) Structural-functional properties of identified excitatory and inhibitory interneurons within pre-Botzinger complex respiratory microcircuits. *J Neurosci* 33:2994–3009.
- Komendantov AO, Ascoli GA (2009) Dendritic excitability and neuronal morphology as determinants of synaptic efficacy. *J Neurophysiol* 101:1847–1866.
- Kong J-H, Fish DR, Rockhill RL, Masland RH (2005) Diversity of ganglion cells in the mouse retina: unsupervised morphological classification and its limits. *J Comp Neurol* 489:293–310.
- Krichmar JL *et al.* (2002) Effects of dendritic morphology on CA3 pyramidal cell electrophysiology: a simulation study. *Brain Res* 941:11–28.
- Kubota Y *et al.* (2011) Selective coexpression of multiple chemical markers defines discrete populations of neocortical GABAergic neurons. *Cereb Cortex* 21:1803–1817.
- Laughlin SB, de Ruyter van Steveninck RR, Anderson JC (1998) The metabolic cost of neural information. *Nat Neurosci* 1:36–41.
- Lee S, Stevens CF (2007) General design principle for scalable neural circuits in a vertebrate retina. *Proc Natl Acad Sci U S A* 104:12931–12935.
- Lewis TL, Mao T, Svoboda K, Arnold DB (2009) Myosin-dependent targeting of transmembrane proteins to neuronal dendrites. *Nat Neurosci* 12:568–576.
- Li Y, Brewer D, Burke RE, Ascoli GA (2005) Developmental changes in spinal motoneuron dendrites in neonatal mice. *J Comp Neurol* 483:304–317.

- Lin B, Masland RH (2005) Synaptic contacts between an identified type of ON cone bipolar cell and ganglion cells in the mouse retina. *Eur J Neurosci* 21:1257–1270.
- Liu G (2004) Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. *Nat Neurosci* 7:373–379.
- Liu Y (2011) The DIADEM and beyond. *Neuroinformatics* 9:99–102.
- Livet J *et al.* (2007) Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 450:56–62.
- London M, Häusser M (2005) Dendritic computation. *Annu Rev Neurosci* 28:503–532.
- Losavio BE *et al.* (2008) Live neuron morphology automatically reconstructed from multiphoton and confocal imaging data. *J Neurophysiol* 100:2422–2429.
- Losonczy A, Makara JK, Magee JC (2008) Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature* 452:436–441.
- Lu J, Tapia J, White O, Lichtman J (2009) The Interscutularis Muscle Connectome. *PLoS Biol* 7:265–277.
- Luczak A (2006) Spatial embedding of neuronal trees modeled by diffusive growth. *J Neurosci Methods* 157:132–141.
- Luisi J, Narayanaswamy A, Galbreath Z, Roysam B (2011) The FARSIGHT trace editor: an open source tool for 3-D inspection and efficient pattern analysis aided editing of automated neuronal reconstructions. *Neuroinformatics* 9:305–315.
- Luo L (2000) RHO GTPASES in neuronal morphogenesis. *Nat Rev Neurosci* 1:173–180.
- Luo L, O’Leary DDM (2005) Axon retraction and degeneration in development and disease. *Annu Rev Neurosci* 28:127–156.
- Macias MY *et al.* (2000) Directed and enhanced neurite growth with pulsed magnetic field stimulation. *Bioelectromagnetics* 21:272–286.
- MacLean JN, Watson BO, Aaron GB, Yuste R (2005) Internal dynamics determine the cortical response to thalamic stimulation. *Neuron* 48:811–823.
- Magariños AM, McEwen BS, Flügge G, Fuchs E (1996) Chronic psychosocial stress causes apical dendritic atrophy of hippocampal CA3 pyramidal neurons in subordinate tree shrews. *J Neurosci* 16:3534–3540.
- Mainen ZF, Sejnowski TJ (1996) Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature* 382:363–366.
- Maletic-Savatic M, Malinow R, Svoboda K (1999) Rapid Dendritic Morphogenesis in CA1 Hippocampal Dendrites Induced by Synaptic Activity. *Science* (80-) 283:1923–1927.
- Malun D, Brunjes PC (1996) Development of olfactory glomeruli: temporal and spatial interactions between olfactory receptor axons and mitral cells in opossums and rats. *J Comp Neurol* 368:1–16.
- Markram H *et al.* (2004) Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* 5:793–807.
- Marks WB, Burke RE (2007) Simulation of motoneuron morphology in three dimensions. I. Building individual dendritic trees. *J Comp Neurol* 503:685–700.
- Martyn T (1793) The language of botany□: being a dictionary of the terms made use of in that science, principally by Linneus: with familiar explanations, and an attempt to establish significant English terms., 2nd editio. London: B. and J. White.

- Matsuki T *et al.* (2010) Reelin and stk25 have opposing roles in neuronal polarization and dendritic Golgi deployment. *Cell* 143:826–836.
- Matyash V, Kettenmann H (2010) Heterogeneity in astrocyte morphology and physiology. *Brain Res Rev* 63:2–10.
- McAllister A (2000) Cellular and Molecular Mechanisms of Dendrite Growth. *Cereb Cortex* 10:963–973.
- McAllister AK, Katz LC, Lo DC (1997) Opposing Roles for Endogenous BDNF and NT-3 in Regulating Cortical Dendritic Growth. *Neuron* 18:767–778.
- McAllister AK, Lo DC, Katz LC (1995) Neurotrophins regulate dendritic growth in developing visual cortex. *Neuron* 15:791–803.
- McGarry LM *et al.* (2010) Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Front Neural Circuits* 4:12.
- Megias M, Emri Z, Freund TF, Gulyás AI (2001) Total number and distribution of inhibitory and excitatory synapses on hippocampal CA1 pyramidal cells. *Neuroscience* 102:527–540.
- Mel BW (1994) Information Processing in Dendritic Trees. *Neural Comput* 6:1031–1085.
- Memelli H, Torben-Nielsen B, Kozloski J (2013) Self-referential forces are sufficient to explain different dendritic morphologies. *Front Neuroinform* 7:1.
- Meyer-Luehmann M *et al.* (2008) Rapid appearance and local toxicity of amyloid-beta plaques in a mouse model of Alzheimer's disease. *Nature* 451:720–724.
- Michaelsen K *et al.* (2010) Fine-tuning of neuronal architecture requires two profilin isoforms. *Proc Natl Acad Sci U S A* 107:15780–15785.
- Migliore M (1996) Modeling the attenuation and failure of action potentials in the dendrites of hippocampal neurons. *Biophys J* 71:2394–2403.
- Migliore M, Ferrante M, Ascoli GA (2005) Signal propagation in oblique dendrites of CA1 pyramidal cells. *J Neurophysiol* 94:4145–4155.
- Miller FD, Kaplan DR (2003) Signaling mechanisms underlying dendrite formation. *Curr Opin Neurobiol* 13:391–398.
- Milo R *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science* (80-) 298:824–827.
- Mirsky JS *et al.* (1998) Database tools for integrating and searching membrane property data correlated with neuronal morphology. *J Neurosci Methods* 82:105–121.
- Mize RR (1984) Computer applications in cell and neurobiology: a review. *Int Rev Cytol* 90:83–124.
- Mottini A, Descombes X, Besse F (2014) From Curves to Trees: A Tree-like Shapes Distance Using the Elastic Shape Analysis Framework. *Neuroinformatics*.
- Myatt DR, Hadlington T, Ascoli GA, Nasuto SJ (2012) Neuromantic - from semi-manual to semi-automatic reconstruction of neuron morphology. *Front Neuroinform* 6:4.
- Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443–453.
- Nikolenko V, Poskanzer KE, Yuste R (2007) Two-photon photostimulation and imaging of neural circuits. *Nat Methods* 4:943–950.
- Nikolić M (2008) The Pak1 Kinase: An Important Regulator of Neuronal Morphology and Function in the Developing Forebrain. *Mol Neurobiol* 37:187–202.

- Nilsson M *et al.* (1999) Enriched environment increases neurogenesis in the adult rat dentate gyrus and improves spatial memory. *J Neurobiol* 39:569–578.
- Núñez-Abades PA, He F, Barrionuevo G, Cameron WE (1994) Morphology of developing rat genioglossal motoneurons studied in vitro: changes in length, branching pattern, and spatial distribution of dendrites. *J Comp Neurol* 339:401–420.
- O’Rourke NA, Weiler NC, Micheva KD, Smith SJ (2012) Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nat Rev Neurosci* 13:365–379.
- Oberlaender M *et al.* (2012) Cell type-specific three-dimensional structure of thalamocortical circuits in a column of rat vibrissa cortex. *Cereb Cortex* 22:2375–2391.
- Overdijk J, Uylings HB, Kuypers K, Kamstra W (1978) An economical, semi-automatic system for measuring cellular tree structures in three dimensions, with special emphasis on Golgi-impregnated neurons. *J Microsc* 114:271–284.
- Parekh R, Armañanzas R, Ascoli GA (2015) The importance of metadata to assess information content in digital reconstructions of neuronal morphology. *Cell Tissue Res* 360:121–127.
- Parekh R, Ascoli GA (2013) Neuronal morphology goes digital: a research hub for cellular and system neuroscience. *Neuron* 77:1017–1038.
- Parrish JZ, Emoto K, Kim MD, Jan YN (2007) Mechanisms that Regulate Establishment, Maintenance, and Remodeling of Dendritic Fields. *Annu Rev Neurosci* 30:399–423.
- Pedrotti B, Colombo R, Islam K (1994) Microtubule associated protein MAP1A is an actin-binding and crosslinking protein. *Cell Motil Cytoskeleton* 29:110–116.
- Peng H *et al.* (2010a) V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat Biotechnol* 28:348–353.
- Peng H, Long F, Zhao T, Myers E (2011) Proof-editing is the bottleneck of 3D neuron reconstruction: the problem and solutions. *Neuroinformatics* 9:103–105.
- Peng H, Ruan Z, Atasoy D, Sternson S (2010b) Automatic reconstruction of 3D neuron structures using a graph-augmented deformable model. *Bioinformatics* 26:i38–i46.
- Pierce JP, Milner TA (2001) Parallel increases in the synaptic and surface areas of mossy fiber terminals following seizure induction. *Synapse* 39:249–256.
- Polavaram S, Gillette TA, Parekh R, Ascoli GA (2014) Statistical analysis and data mining of digital reconstructions of dendritic morphologies. *Front Neuroanat* 8.
- Polleux F, Morrow T, Ghosh A (2000) Semaphorin 3A is a chemoattractant for cortical apical dendrites. *Nature* 404:567–573.
- Pyapali GK *et al.* (1998) Dendritic properties of hippocampal CA1 pyramidal neurons in the rat: Intracellular staining in vivo and in vitro. *J Comp Neurol* 391:335–352.
- Quinlan EM, Halpain S (1996) Emergence of Activity-Dependent, Bidirectional Control of Microtubule-Associated Protein MAP2 Phosphorylation during Postnatal Development. *J Neurosci* 16:7627–7637.
- Radley JJ *et al.* (2004) Chronic behavioral stress induces apical dendritic reorganization in pyramidal neurons of the medial prefrontal cortex. *Neuroscience* 125:1–6.

- Rajnicek A, Gow N, McCaig C (1992) Electric field-induced orientation of rat hippocampal neurones in vitro. *Exp Physiol* 77:229–232.
- Rall W (1969) Time Constants and Electrotonic Length of Membrane Cylinders and Neurons. *Biophys J* 9:1483–1508.
- Rand WM (1971) Objective Criteria for the Evaluation of Clustering Methods. 66:846–850.
- Rapp M, Segev I, Yarom Y (1994) Physiology, morphology and detailed passive models of guinea-pig cerebellar Purkinje cells. *J Physiol* 474:101–118.
- Ristanović D, Milosević NT, Stulić V (2006) Application of modified Sholl analysis to neuronal dendritic arborization of the cat spinal cord. *J Neurosci Methods* 158:212–218.
- Rocchi MBL, Sisti D, Albertini MC, Teodori L (2007) Current trends in shape and texture analysis in neurology: aspects of the morphological substrate of volume and wiring transmission. *Brain Res Rev* 55:97–107.
- Rodriguez A, Ehlenberger DB, Hof PR, Wearne SL (2009) Three-dimensional neuron tracing by voxel scooping. *J Neurosci Methods* 184:169–175.
- Ropireddy D *et al.* (2011) Axonal morphometry of hippocampal pyramidal neurons semi-automatically reconstructed after in vivo labeling in different CA3 locations. *Brain Struct Funct* 216:1–15.
- Ropireddy D, Ascoli GA (2011) Potential Synaptic Connectivity of Different Neurons onto Pyramidal Cells in a 3D Reconstruction of the Rat Hippocampus. *Front Neuroinform* 5:5.
- Ruffalo M, LaFramboise T, Koyutürk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27:2790–2796.
- Sagasti A, Guido MR, Raible DW, Schier AF (2005) Repulsive interactions shape the morphologies and functional arrangement of zebrafish peripheral sensory arbors. *Curr Biol* 15:804–814.
- Samsonovich A V., Ascoli GA (2003) Statistical morphological analysis of hippocampal principal neurons indicates cell-specific repulsion of dendrites from their own cell. *J Neurosci Res* 71:173–187.
- Samsonovich A V., Ascoli GA (2005) Statistical determinants of dendritic morphology in hippocampal pyramidal neurons: A hidden Markov model. *Hippocampus* 15:166–183.
- Samsonovich A V., Ascoli GA (2006) Morphological homeostasis in cortical dendrites. *Proc Natl Acad Sci U S A* 103:1569–1574.
- Santana R *et al.* (2013) Classification of neocortical interneurons using affinity propagation. *Front Neural Circuits* 7:185.
- Scelfo B, Strata P (2005) Correlation between multiple climbing fibre regression and parallel fibre response development in the postnatal mouse cerebellum. *Eur J Neurosci* 21:971–978.
- Schaap M *et al.* (2009) Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. *Med Image Anal* 13:701–714.

- Schaefer AT, Larkum ME, Sakmann B, Roth A (2003) Coincidence detection in pyramidal neurons is tuned by their dendritic branching pattern. *J Neurophysiol* 89:3143–3154.
- Schmitz SK *et al.* (2011) Automated analysis of neuronal morphology, synapse number and synaptic recruitment. *J Neurosci Methods* 195:185–193.
- Schneider CJ, Bezaire M, Soltesz I (2012) Toward a full-scale computational model of the rat dentate gyrus. *Front Neural Circuits* 6:83.
- Scorcioni R, Ascoli GA (2005) Algorithmic reconstruction of complete axonal arborizations in rat hippocampal neurons. *Neurocomputing* 65-66:15–22.
- Scorcioni R, Lazarewicz MT, Ascoli GA (2004) Quantitative morphometry of hippocampal pyramidal cells: Differences between anatomical classes and reconstructing laboratories. *J Comp Neurol* 473:177–193.
- Scorcioni R, Polavaram S, Ascoli GA (2008) L-Measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. *Nat Protoc* 3:866–876.
- Scott EK, Luo L (2001) How do dendrites take their shape? *Nat Neurosci* 4:359–365.
- Segura V, Ouangraoua A, Ferraro P, Costes E (2008) Comparison of tree architecture using tree edit distances: application to 2-year-old apple hybrids. *Euphytica* 161:155–164.
- Senft SL (2011) A brief history of neuronal reconstruction. *Neuroinformatics* 9:119–128.
- Shepherd GMG, Svoboda K (2005) Laminar and Columnar Organization of Ascending Excitatory Projections to Layer 2/3 Pyramidal Neurons in Rat Barrel Cortex. *J Neurosci* 25:5670–5679.
- Sholl DA (1953) Dendritic organization in the neurons of the visual and motor cortices of the cat. *J Anat* 87:387–406.1.
- Sieck GC, Prakash YS (1995) Fatigue at the neuromuscular junction. Branch point vs. presynaptic vs. postsynaptic mechanisms. *Adv Exp Med Biol* 384:83–100.
- Sin WC, Haas K, Ruthazer ES, Cline HT (2002) Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* 419:475–480.
- Smit-Rigter LA *et al.* (2012) Prenatal fluoxetine exposure induces life-long serotonin 5-HT3 receptor-dependent cortical abnormalities and anxiety-like behaviour. *Neuropharmacology* 62:865–870.
- Snider J, Pillai A, Stevens CF (2010) A universal property of axonal and dendritic arbors. *Neuron* 66:45–56.
- Soloway AS, Pucak ML, Melchitzky DS, Lewis DA (2002) Dendritic morphology of callosal and ipsilateral projection neurons in monkey prefrontal cortex. *Neuroscience* 109:461–471.
- Sporns O, Chialvo DR, Kaiser M, Hilgetag CC (2004) Organization, development and function of complex brain networks. *Trends Cogn Sci* 8:418–425.
- Spratling MW (2002) Cortical region interactions and the functional role of apical dendrites. *Behav Cogn Neurosci Rev* 1:219–228.
- Spruston N (2008) Pyramidal neurons: dendritic structure and synaptic integration. *Nat Rev Neurosci* 9:206–221.

- Spruston N, Schiller Y, Stuart G, Sakmann B (1995) Activity-dependent action potential invasion and calcium influx into hippocampal CA1 dendrites. *Science* (80-) 268:297–300.
- Staiger JF *et al.* (2004) Functional diversity of layer IV spiny neurons in rat somatosensory cortex: quantitative morphology of electrophysiologically characterized and biocytin labeled cells. *Cereb cortex* 14:690–701.
- Steinhaus H (1956) Sur la division des corp materiels en parties. *Bull l'Académie Pol des Sci Cl 3* 1:801–804.
- Stepanyants A, Chklovskii D (2005) Neurogeometry and potential synaptic connectivity. *Trends Neurosci* 28:387–394.
- Stepanyants A, Tamás G, Chklovskii DB (2004) Class-Specific Features of Neuronal Wiring. *Neuron* 43:251–259.
- Stettler DD *et al.* (2006) Axons and synaptic boutons are highly dynamic in adult visual cortex. *Neuron* 49:877–887.
- Stiefel KM, Sejnowski TJ (2007) Mapping Function Onto Neuronal Morphology. *J Neurophysiol* 98:513–526.
- Sugihara I, Wu H, Shinoda Y (1996) Morphology of axon collaterals of single climbing fibers in the deep cerebellar nuclei of the rat. *Neurosci Lett* 217:33–36.
- Sulkowski MJ *et al.* (2011) Turtle functions downstream of Cut in differentially regulating class specific dendrite morphogenesis in Drosophila. *PLoS One* 6:e22611.
- Szebenyi G, Callaway JL, Dent EW, Kalil K (1998) Interstitial Branches Develop from Active Regions of the Axon Demarcated by the Primary Growth Cone during Pausing Behaviors. *J Neurosci* 18:7930–7940.
- Tamamaki N, Nojyo Y (1990) Disposition of the slab-like modules formed by axon branches originating from single CA1 pyramidal neurons in the rat hippocampus. *J Comp Neurol* 291:509–519.
- Tamamaki N, Nojyo Y (1991) Crossing fiber arrays in the rat hippocampus as demonstrated by three-dimensional reconstruction. *J Comp Neurol* 303:435–442.
- Tamori Y (1993) Theory of dendritic morphology. *Phys Rev E* 48:3124.
- Teeter CM, Stevens CF (2011) A General Principle of Neural Arbor Branch Density. *Curr Biol* 21:2105–2108.
- Thompson C *et al.* (2008) Genomic Anatomy of the Hippocampus. *Neuron* 60:1010–1021.
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
- Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 6:e18093.
- Thomson AM, Armstrong WE (2011) Biocytin-labelling and its impact on late 20th century studies of cortical circuitry. *Brain Res Rev* 66:43–53.
- Ting C-Y *et al.* (2014) Photoreceptor-derived activin promotes dendritic termination and restricts the receptive fields of first-order interneurons in Drosophila. *Neuron* 81:830–846.

- Torben-Nielsen B, De Schutter E (2014) Context-aware modeling of neuronal morphologies. *Front Neuroanat* 8:92.
- Turner DA *et al.* (1995) Morphometric and electrical properties of reconstructed hippocampal CA3 neurons recorded in vivo. *J Comp Neurol* 356:580–594.
- Tyrrell JA *et al.* (2007) Robust 3-D modeling of vasculature imagery using superellipsoids. *IEEE Trans Med Imaging* 26:223–237.
- Uemura E, Carriquiry A, Kliemann W, Goodwin J (1995) Mathematical modeling of dendritic growth in vitro. *Brain Res* 671:187–194.
- Uylings H, van Pelt J (2002) Measures for quantifying dendritic arborizations. *Netw Comput Neural Syst* 13:397–414.
- Van Ooyen A (2011) Using theoretical models to analyse neural development. *Nat Rev Neurosci* 12:311–326.
- Van Ooyen A, Duijnhouwer J, Remme M, van Pelt J (2002) The effect of dendritic topology on firing patterns in model neurons. *Netw Comput Neural Syst* 13:311–325.
- Van Pelt J *et al.* (1992) Tree asymmetry—A sensitive and practical measure for binary topological trees. *Bull Math Biol* 54:759–784.
- Van Pelt J (1997) Effect of Pruning on Dendritic Tree Topology. *J Theor Biol* 186:17–32.
- Van Pelt J, Dityatev AE, Uylings HB (1997) Natural variability in the number of dendritic segments: model-based inferences about branching during neurite outgrowth. *J Comp Neurol* 387:325–340.
- Van Pelt J, Verwer R (1983) The exact probabilities of branching patterns under terminal and segmental growth hypotheses. *Bull Math Biol* 45:269–285.
- Van Pelt J, Verwer R (1986) Topological properties of binary trees grown with order-dependent branching probabilities. *Bull Math Biol* 48:197–211.
- Van Pelt J, Verwer RW, Uylings HB (1986) Application of growth models to the topology of neuronal branching patterns. *J Neurosci Methods* 18:153–165.
- Van Praag H, Kempermann G, Gage FH (2000) Neural consequences of environmental enrichment. *Nat Rev Neurosci* 1:191–198.
- Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, Fourth. New York: Springer.
- Verwer R, Van Pelt J (1990) Analysis of binary trees when occasional multifurcations can be considered as aggregates of bifurcations. *Bull Math Biol* 52:629–641.
- Vetter P, Roth A, Häusser M (2001) Propagation of Action Potentials in Dendrites Depends on Dendritic Morphology. *J Neurophysiol* 85:926–937.
- Wan Y *et al.* (2015) BlastNeuron for Automated Comparison, Retrieval and Clustering of 3D Neuron Morphologies. *Neuroinformatics* In Press.
- Wang SS-H *et al.* (2008) Functional trade-offs in white matter axonal scaling. *J Neurosci* 28:4047–4456.
- Wang Y *et al.* (2002) Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cereb Cortex* 12:395–410.
- Wang Y *et al.* (2006) Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat Neurosci* 9:534–542.

- Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* 58:236–244.
- Wearne SL *et al.* (2005) New techniques for imaging, digitization and analysis of three-dimensional neural morphology on multiple scales. *Neuroscience* 136:661–680.
- Weihs C, Ligges U, Luebke K, Raabe N (2005) klaR Analyzing German Business Cycles. In: Data Analysis and Decision Support (Baier D, Decker R, Schmidt-Thieme L, eds), pp 335–343. Berlin: Springer-Verlag.
- Weiler N *et al.* (2008) Top-down laminar organization of the excitatory network in motor cortex. *Nat Neurosci* 11:360–366.
- Wen Q, Chklovskii DB (2008) A Cost-Benefit Analysis of Neuronal Morphology. *J Neurophysiol* 99:2320–2328.
- Whitford KL, Dijkhuizen P, Polleux F, Ghosh A (2002) Molecular control of cortical dendrite development. *Annu Rev Neurosci* 25:127–149.
- Williams SR, Stuart GJ (2000) Action Potential Backpropagation and Somato-dendritic Distribution of Ion Channels in Thalamocortical Neurons. *J Neurosci* 20:1307–1317.
- Wong ROL, Ghosh A (2002) Activity-dependent regulation of dendritic growth and patterning. *Nat Rev Neurosci* 3:803–812.
- Woodruff A, Xu Q, Anderson SA, Yuste R (2009) Depolarizing effect of neocortical chandelier neurons. *Front Neural Circuits* 3:15.
- Wright SN *et al.* (2013) Digital reconstruction and morphometric analysis of human brain arterial vasculature from magnetic resonance angiography. *Neuroimage* 82:170–181.
- Yamamoto H *et al.* (1988) Dephosphorylation of microtubule proteins by brain protein phosphatases 1 and 2A, and its effect on microtubule assembly. *J Neurochem* 50:1614–1623.
- Yang J *et al.* (2015) Pathological Axonal Death through a MAPK Cascade that Triggers a Local Energy Deficit. *Cell* 160:161–176.
- Yates D, Moore DS, Starnes DS (2002) The Practice of Statistics: TI-83/89 Graphing Calculator Enhanced. W. H. Freeman.
- Yoshimura T, Arimura N, Kaibuchi K (2006) Signaling networks in neuronal polarization. *J Neurosci* 26:10626–10630.
- Yu W *et al.* (2000) Depletion of a microtubule-associated motor protein induces the loss of dendritic identity. *J Neurosci* 20:5782–5791.
- Zador A, Agmon-Snir H, Segev I (1995) The morphoelectrotonic transform: a graphical approach to dendritic function. *J Neurosci* 15:1669–1682.
- Zawadzki K *et al.* (2012) Morphological homogeneity of neurons: searching for outlier neuronal cells. *Neuroinformatics* 10:379–389.
- Zhang K (1996) A constrained edit distance between unordered labeled trees. *Algorithmica* 15:205–222.

BIOGRAPHY

Todd Aaron Gillette graduated from Atholton High School, Columbia, Maryland, in 1999. He received his Bachelor of Science from Swarthmore College in 2003 with a double major in Engineering (BS) and Computer Science (BA). He was employed as a software designer at Vivakos, Inc in Alexandria, Virginia for 3 years until starting the Ph.D. program in Neuroscience in September 2006 and becoming a Ph.D. candidate in 2010.