12-2013

Constraints on Patterns of Abundance and Aggregation in Biological Systems

Kenneth J. Locey
Utah State University

CONSTRAINTS ON PATTERNS OF ABUNDANCE AND AGGREGATION IN BIOLOGICAL SYSTEMS

by

Kenneth J. Locey

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Biology

Approved:

_____	_____
Ethan P. White Major Professor	Jeanette Norton Committee Member
_____	_____
S. K. Morgan Ernest Committee Member	Jacob Parnell Committee Member
_____	_____
David Koons Committee Member	Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah
2013

ABSTRACT

Constraints on patterns of abundance and aggregation

in biological systems

by

Kenneth J. Locey, Doctor of Philosophy

Utah State University, 2013

Major Professor: Ethan P. White
Department: Biology

Understanding mechanisms that drive the structure of biological systems is a primary goal of all biological sciences. From the shapes of molecules to the structure of ecological communities, all biological systems are characterized by patterns in the organization of their components (i.e. structure). Despite their complexity, the structure of biological systems is often characterized by relatively simply patterns in properties such as aggregation and abundance. The overarching goal of my research was to explore how patterns in these structural properties may be constrained by general state variables such as the size of a biological system and number of elements, parts, or components within it (e.g. number base pairs in a genome, number of chromosomes, number of species and individual organisms in community).

For example, aggregation of nucleotide base pairs is a structural property that reflects, among other things, the encoding of genetic information. Using over 300

microbial genomes, I demonstrate that aggregation among nucleotides is constrained by genome length (but not chromosome length) and that levels of aggregation consistently differ between regions of coding and noncoding DNA. Likewise, in ecological systems, the distribution of abundance among species is a primary pattern of ecological structure that provides a full characterization of commonness and rarity. Using over 9000 communities of plants, animals, fungi, and microorganisms, I demonstrate how species abundance distributions are constrained by total abundance and species richness, and that the strength of this influence differs between communities of microbes and those of conspicuous plants and animals.

In conclusion, general variables such as the size of a system and the number of elements within it can greatly constrain simple structural properties at scales of biology from individual genomes to ecological communities. The influence of these constaints is poorly understood and this lack of understanding influences our interpretation of processes underlying biological structure. My research provides a step towards a general understanding of the constraining influence of general variables on patterns of aggregation and abundance across biology.

(113 pages)

PUBLIC ABSTRACT

Constraints on patterns of abundance and aggregation

in biological systems

by

Kenneth J. Locey, Doctor of Philosophy

Utah State University, 2013

Major Professor: Ethan P. White
Department: Biology

Understanding the mechanisms that structure biological systems is a primary goal of biology. My research shows that the biological structure is constrained in important ways by general variables such as the number of base pairs in a genome and the number of individuals and species in a community. I used a combination of macroecology, bioinformatics, statistics, mathematics, and advanced computing to pursue my research and published several peer-reviewed scientific manuscripts and open-source software as a result.

I was funded through a combination of fellowships and scholarships awarded by the Utah State University School of Graduate Studies, College of Science, and Department of Biology, as well as teaching assistantships awarded through the Department of Biology at Utah State University, and research assistantships funded through a CAREER grant from the U.S. National Science Foundation (DEB-0953694)

awarded to my advisor, Dr. Ethan White. With the help of my advisor, I also obtained a computing grant from Amazon Web Services in the amount of $7,500. Altogether, funding for my research and education totaled approximately $123,500.

Using over 9000 communities of plants, animals, fungi, and microorganisms, I demonstrated that the forms of empirical species abundance distributions (SADs) are constrained by total abundance and species richness. Using over 300 microbial genomes, I demonstrate that nucleotide aggregation is constrained by genome length and differs between regions of coding and noncoding DNA. General state variables of genomes and ecological communities (i.e. genome length, total abundance and species richness) constrain simple structural properties of each system.

ACKNOWLEDGMENTS

I wish to thank my committee members, Morgan Ernest, David Koons, Jeanette Norton, and Jacob Parnell, for their encouragement and support of my doctoral research and preparation of this dissertation. Their insights, questions, and varied expertise were invaluable. I also wish to thank my advisor, Ethan White, without whose encouragement, mentoring, patience, and support this dissertation work would not have been possible. Ethan provided tremendous creative latitude and freedom to explore disparate areas of biology and ecology, and unfailingly provided advice that always proved to be useful if not absolutely correct. Thanks to the members of Weecology: Ethan White, Morgan Ernest, Xiao Xiao, Dan McGlinn, Zach Brym, Glenda Yenni, Benjamin Morris, Sarah Supp, Elita Baldridge, Kate Thibault, Erica Christensen, and Kristina Riemer. I am grateful to each of you for your help, discussions, friendship, and the time spent learning from and alongside each of you.

I also wish to thank the numerous individuals involved in collecting and providing the data used in this research, including the essential citizen scientists who collect data for the North American Breeding Birds Survey and the Christmas Bird Count, USGS and CWS scientists and managers, researchers who collected and sequenced the environmental genomic data from microbial communities, the MG-RAST project, the Audubon Society, the U.S. Forest Service, the Missouri Botanical Garden, and Alwyn H. Gentry. I thank Xiao Xiao, Daniel McGlinn, Jay T. Lennon, Paul A. Stone, Paul Cliften, Bill Burnside, Justin Kitzes, Jacob Parnell, James O'Dwyer and anonymous reviewers for fruitful discussions and critical comments of my research. During my tenure at USU, my

research was supported by a CAREER grant from the U.S. National Science Foundation to EPW (DEB-0953694) awarded to Ethan White, by a research grant from Amazon Web Services to Ethan White and myself, and through graduate fellowships and scholarships awarded through the School of Graduate Studies, College of Science, and Department of Biology at Utah State University.

Most of all, I wish to thank my wife, Lisa. Her unconditional love and support allowed me to pursue my dissertation work with full and unrestricted fervor. For more than five years, I paid little heed to the hour of the day and day of the week for the sake of unrelentingly pursuing science on several disparate fronts. This, of course, was only possible because of Lisa's hard work, unfailing support, and understanding.

Kenneth J. Locey

CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Appendices

CHAPTER 1

INTRODUCTION

Understanding mechanisms that influence the structure of biological systems is a primary goal of all biological sciences (Watson & Crick 1953, Paine 1974, Gehring 1987, Brown 1995). From the structure of molecules involved in gene expression and metabolic pathways, to the structure of populations and communities characterized by patterns of abundance among alleles and species, all biological systems are characterized by patterns in the organization of their components (i.e. structure). Despite their inherent complexity, at least two structural properties apply to most any biological system, i.e. aggregation (i.e. properties of grouping and clustering) and abundance. These structural properties may, in turn, be constrained by general variables that reflect the size and number of elements in a biological system (e.g. genome length, length and number of chromosomes, total community abundance, number of species) (Almirantis & Provata 1997, Bohlin et al. 2009, McGill 2010, Harte 2011).

At the molecular scale, the number of nucleotide base pairs is a basic attribute of any genome. Likewise, the aggregation of nucleotides is a structural property that reflects the encoding of genetic information and likely differs between regions of DNA that code and do not code for proteins (Almirantis & Provata 1997, Bohlin et al. 2009). It has also been suggested that forces related to entropy may drive aggregation within genomes and the self-assembly of genomic structures (Marenduzzo et al. 2006).

At the macroecological scale, total abundance and species richness are basic attributes of ecological communities that inherently influence the distribution of abundance among species; a primary pattern of ecological structure (McGill et al. 2007). Recently, it has been shown that models primarily constrained by total abundance and species richness can explain the majority of variation in abundances among species (Harte et al. 2008, Harte 2011, White et al. 2012). Indeed, species richness and total abundance constitute two primary state variables in the Maximum Entropy Theory of Ecology (Harte 2011).

The constraining influence of general variables on structural patterns of abundance and aggregation has rarely been studied across disparate fields of biology or even within a single context. The overarching goal of my research was to show how the structural properties of aggregation and abundance are constrained across biological systems by general variables that relate to the size of a system and its number of primary components (e.g. genome length, chromosome length and number, total community abundance, species richness). I demonstrate the breadth of the constraining influence of these general variables on abundance and aggregation by focusing on the molecular and ecological scales, i.e., aggregation of nucleotides in genomes as constrained by genome length and chromosome number, and distributions of abundance in ecological communities as constrained by total abundance and species richness.

References

Almirantis Y. & Provata A. (1997). The “clustered structure” of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. *B. Math. Biol.*, 59, 975-992.

Bohlin J., Hardy S.P. & Ussery, D.W. (2009). Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes. *BMC Genomics*, 10, 346.

Brown, J.H. (1995). *Macroecology*. Univ. Chicago Press, Chicago.

Gehring, W.J. (1987). Homeo boxes in the study of development. *Science*, 236, 1245-1252.

Harte, J. (2011). *Maximum Entropy and Ecology*. Oxford Univ. Press, Oxford.

Harte, J., Zillio, T., Conlisk, E. & Smith, A.B. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology*, 89, 2700-2711.

Marenduzzo D., Micheletti C. & Cook P.R. (2006). Entropy-drive genome organization. *Biophys. J.*, 90, 3712-3721.

McGill, B.J. (2010). Towards a unification of unified theories of biodiversity. *Ecol. Lett.*, 13, 627-642.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K. *et al*. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, 995–1015.

Paine, R.T. (1974). Intertidal community structure: Experimental studies on the relationship between a dominant competitor and its principal predator. *Oecologia*, 15, 93-120.

Watson J.D. & Crick F.H. (April 1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid" (PDF). *Nature*, 171, 737-738.

White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772-1778.

CHAPTER 2

HOW SPECIES RICHNESS AND TOTAL ABUNDANCE CONSTRAIN THE DISTRIBUTION OF ABUNDANCE*

Abstract

The species abundance distribution (SAD) is one of the most intensively studied distributions in ecology and its hollow-curve shape is one of ecology's most general patterns. We examine the SAD in the context of all possible forms having the same richness (S) and total abundance (N), i.e. the feasible set. We find that feasible sets are dominated by similarly-shaped hollow-curves, most of which are highly correlated with empirical SADs (most R^2 values > 75%), revealing a strong influence of N and S on the form of the SAD and an *a priori* explanation for the ubiquitous hollow-curve. Empirical SADs are often more hollow and less variable than the majority of the feasible set, revealing exceptional unevenness and relatively low natural variability among ecological communities. We discuss the importance of the feasible set in understanding how general constraints determine observable variation and influence the forms of predicted and empirical patterns.

*Coauthored by: Locey, K.J. & White, E.P. (2013). How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, 16, 1177-1185.

Introduction

The species abundance distribution (SAD) is one of the most widely studied patterns in ecology and exhibits a consistent structure with many rare and few common species; the canonical "hollow curve" (McGill et al. 2007). The form of the SAD has been predicted by a variety of models based on an array of different processes including niche differentiation (e.g. Sugihara 1980), stochastic population dynamics (e.g. Hubbell 2001), and the structure of abundance across a species range (e.g. McGill & Collins 2003). Though SADs are potentially influenced by some or all of these processes, the ability to distinguish between different structuring processes depends on the presence of sufficient variation among the possible shapes of the SAD (Haegeman & Loreau 2008). If most of the possible SADs have similar shapes, it will be difficult to determine what processes generated them.

Haegeman & Loreau (2008) introduced the use of the set of all possible distributions (the feasible set) to examine ecological patterns and theory. They argue that if the feasible set is small then there is little information in the pattern being examined. Likewise, if theoretical predictions do not deviate from the center of the feasible set, then they may provide limited information about process. To explore the implications of these ideas for understanding the species abundance distribution we use McGill et al.'s (2007) definition of the SAD as the "vector of abundances of all species present in a community". This distribution is necessarily influenced by two values: total abundance (N; i.e. the number of individuals in a community) and species richness (S; i.e. the number of species in a community). Though ecological theories often use N and S as inputs to fit or predict the shape of the SAD (McGill 2010), knowing N and S constrains

the form of the SAD in ways that ecologists rarely address. Specifically, there are a limited number of ways that the abundances of S species can sum to a total abundance of N, and thus, there is a limited feasible set of uniquely-shaped SADs for any combination of N and S. For example, there is only one possible SAD form when S = 1 and N = 1 (i.e. {1}) and only two possible forms if S = 2 and N = 5 (i.e. {4,1}, {3,2}). As we show, N and S not only determine the number of possible SADs but also the general form of the possible distributions, making it necessary to understand how the properties of the feasible set constrain the form of the SAD and how constraints such as N and S influence empirical patterns and the predictions of ecological models.

We refer to each uniquely-shaped SAD within the feasible set as a macrostate (i.e. an unordered vector of unlabeled species abundances). This differs from a microstate, which refers to a unique distribution of individuals among species leading to a specific macrostate. The terms feasible set, macrostate, and microstate have been used in recent applications of entropy maximization (MaxEnt) to macroecology (Haegeman & Loreau 2008, McGill & Nekola 2010; Harte 2011). In short, MaxEnt infers the most likely macrostate as that with the most microstates based on sets of state variables (e.g. N, S) and related constraints. Though the framework of MaxEnt implies the existence of a feasible set, MaxEnt does not explicitly consider it. Here, we focus solely on the distribution of macrostates within the feasible set without considering the numbers of ways in which macrostates can arise. As we show, feasible sets have strong central tendencies, meaning that most of the possible macrostates have similar shapes. If empirical SADs have shapes similar to those near the center of the feasible set, then there

may be little ecological information in the shape of the SAD beyond that contained in N and S. Since most of the observable forms of the abundance distribution have shapes that are very similar to this central tendency, many different processes will result in distributions of the same general form as will many different models. This observation goes beyond the issue of equivalent models (e.g., Pielou 1975, McGill et al. 2007), suggesting that many different models and empirical patterns may be expected to take similar forms because most possible states of the pattern are similar in shape (White et al. 2012). However, if the shape of an SAD is exceptional to the majority of shapes with the same N and S, then this exceptional evenness or unevenness would require an explanation, especially if consistent across communities. Consequently, the feasible set provides a context for understanding whether predicted and empirical patterns are exceptional to or representative of the majority of possible forms. As such, studies of the SAD would benefit from considering the shape of the SAD relative to the feasible set, rather than the shape of the SAD *per se*.

Here, we explore general properties of the feasible set and reveal the strength of the influence of N and S on the shape of the SAD. We use the feasible set as a contextual framework for understanding how richness and abundance necessarily constrain ecological patterns. We show that most of the possible SAD shapes are similarly-shaped hollow-curves, revealing an *a priori* reason for the ubiquitous hollow-curve. Using one of the most taxonomically diverse and geographically expansive data compilations in community ecology, we show that the central tendency of the feasible set is strongly correlated with empirical SAD patterns within and among sites for birds, mammals, trees,

and metagenomic datasets of prokaryotes and fungi. Moving beyond single SADs, we use ensembles of SADs with the same values of N and S to assess relationships between the variance predicted by the feasible set and that observed in ecological systems. We discuss the importance of using the feasible set as a context for understanding variation in the forms of empirical patterns and the inference that can be drawn from models that successfully predict them.

Methods

Finding macrostates of feasible sets

Finding all possible macrostates for a community of a particular total abundance (N) and species richness (S) is equivalent to finding all unordered ways of summing S positive integers to obtain the positive integer N, a combinatorial approach known as integer partitioning (Andrews & Eriksson 2004). For example, the feasible set for N = 10 and S = 3 is {8+1+1, 7+2+1, 6+3+1, 6+2+2, 5+4+1, 5+3+2, 4+4+2, 4+3+3}. Different unordered sets of S integers that sum to N are partitions (i.e. macrostates) of N and S. Hence, sets of the same integers in different order, e.g. {8+1+1, 1+8+1, 1+1+8}, constitute the same partition. Likewise, each would produce the same frequency distribution (i.e. two 1's, one 8) and the same rank distribution (i.e. 8+1+1). Several algorithms are available for integer partitioning problems, such as finding the size of the feasible set for a given N and S (Nijenhuis & Wilf 1978). We used the implementation of these algorithms in the free open-source Python-based Sage computer algebra system (http://www.sagemath.org/).

Generating the feasible set for a community of a particular N and S can require a large amount of time and computational memory. This is because feasible sets become large for communities of realistic size; a result of combinatorial explosion (i.e. large changes in the number of possible outcomes for small changes in the values of inputs). For example, there are nearly 8.8×10^{14} macrostates for N = 1,000 and S = 10 and nearly 6.28×10^{26} macrostates for N = 1,000 and S = 50. While complete enumeration of the feasible set can be untenable for many values of N and S, the form of the feasible set space can be determined by randomly sampling macrostates from the feasible set. We used the random partition algorithm described by Nijenhuis & Wilf (1978) and implemented in Sage to generate uniform random samples of feasible sets.

The partitioning algorithm we used, and all currently implemented integer partitioning algorithms, generates random partitions (i.e. macrostates) based on N but not S. We randomly drew partitions of N and rejected partitions that did not have S parts; an approach that can be computationally expensive. For example, randomly drawing one macrostate for N = 1,000 and S = 10 requires drawing from a feasible set of nearly 2.4×10^{31} macrostates, one of the roughly 8.9×10^{14} for which S = 10; a probability of nearly 3.7×10^{-17}. Consequently, we used substantial computational resources (one in-house cluster of three dual Quad Core Intel Xeon 3 GHz processors with 16 GB of RAM each, plus 20 High-CPU Extra Large Amazon Web Service instances with 7GB of RAM each and a total of 160 AWS cores) and computational time (> 10,000 compute hours) to generate random macrostates for combinations of N and S. Code for replicating our

analyses are available at https://github.com/weecology/feasiblesets. All software required to run our scripts (e.g. Sage, Numpy, Python) is free and open source.

We chose the integer partitioning approach to the feasible set over the random walk method used by Haegeman & Loreau (2008) because it is conceptually simpler and, by definition, yields uniform random samples of the feasible set without requiring decisions regarding burn-in periods and the number of steps between samples. However, this approach can be very slow for some combinations of N and S, and further research comparing the speed and accuracy of these two approaches would be valuable.

Data

We used a subset of previously compiled datasets of site-specific species abundance data (see White et al. 2012). Our subset represents 9,562 different sites of bird, tree, and mammal communities. The data set includes four continental-to-global scale surveys, including the Christmas Bird Count (129 sites) (CBC; National Audubon Society 2002), North American Breeding Bird Survey (1,586 sites) (BBS; Sauer et al. 2011), Gentry's Forest Transect Data Set (182 sites) (GENTRY; Phillips & Miller 2002), Forest Inventory Analysis (7,359 sites) (FIA; U.S. Department of Agriculture 2010), and one global-scale data compilation, the Mammal Community Database (42 sites) (MCDB; Thibault et al. 2011). White et al. (2012) used one year of sampling for each site and only used data for communities with a minimum of 10 species (Ulrich et al. 2010). We included only sites with combinations of N and S for which random macrostates could be generated based on reasonable computational effort. This includes large fractions of all of the datasets except for CBC, which only includes ~6% of the original sites. Additionally,

we restricted our analysis of FIA to natural forest stands (e.g. absence of human disturbance, plots without artificial regeneration, plots without silvicuture treatment). More details regarding the data can be found in Appendix C and in Appendix A of White et al. (2012).

We also compiled relative abundance data at the species level from five microbial metagenome projects for a total of 264 surveys of geographically distinct bacterial, archaeal, and indoor fungal communities. Metagenomes are produced from genetic material recovered from environmental samples and are the primary means of studying microbial diversity *in situ*. Despite the lack of a universally accepted microbial species definition, there is a well-established convention for demarcating species-level units. Taxonomic levels representing species are commonly delineated at 97% 16S rRNA sequence similarity for prokaryotes and 97% rRNA sequence and rRNA related ITS (internal transcribed spacer) sequence similarity for fungi (Roselló-Mora & Amann 2001, Schloss & Handelsman 2006, Marshal et al. 2008, Amend et al. 2010, Chu et al. 2010, Flores et al. 2011, Fierer et al. 2012). This convention was used by studies that generated the metagenomic data used in our study.

We used SAD data from region-to-global scale PCR-targeted projects from the metagenomics server MG-RAST (Meyer et al. 2008). PCR-targeted (i.e. amplicon sequenced) approaches provide better overall coverage of a specific gene (e.g. 16S rRNA) than a random shotgun approach by sequencing an amplified target gene. We used the rRNA library provided by MG-RAST (i.e. M5RNA) to obtain SAD data for each metagenome used in our study. We used common thresholds for sequence comparison

and species-level determination (Lazarevic et al. 2009; Lamendella et al. 2011) including a maximum e-value (probability of observing an equal or better match in a database of a given size) cutoff of $1e^{-5}$, a minimum alignment length of 50 base pairs, and a minimum percent identity of 97% to the M5RNA reference sequence. However, because microbial species are sometimes defined below or above 97% (e.g. Webster et al. 2010, Martiny et al. 2011) we also analyzed microbial communities at 95% and 99% species-level cutoffs.

We compiled metagenomic data into datasets representing aquatic prokaryotic communities (48 metagenomes), terrestrial prokaryotic communities (92 metagenomes), and terrestrial fungal communities (124 metagenomes). The aquatic datasets (AQUA) included the Archaeal and Bacterial Diversity of Geographically and Geologically Distinct Deep-Sea Hydrothermal Vent Mineral Deposits project (Flores et al. 2011) and the Catlin Arctic Survey of bacterial and archaeal diversity (www.catlin.com/en/Responsibility/CatlinArcticSurvey). The terrestrial prokaryotic datasets (TERA) included the archaeal and bacterial diversity of the Lauber 88 Soils project (Fierer et al. 2012), and the Chu Arctic Soils project (Chu et al. 2010). The terrestrial fungi dataset was a global-scale survey of fungal community data sampled from indoor habitats of human cities (Amend et al. 2010). Detailed information about each metagenome project is available on the MG-RAST website (http://metagenomics.anl.gov/) and additional details on our use of microbial metagenomes is available in Appendix C.

Our compilation of data is taxonomically diverse. As such, there are differences among our datasets that should be recognized. First, despite our use of accepted species

level delineations for microbes, these species and communities do not represent the same ecological and evolutionarily meaningful units as our other datasets, i.e., genetically distinct populations of biological species. Whereas our macrobial data represent a few well-known members of one domain (i.e. Eukaryota), our microbial datasets include many poorly understood members from all three. Second, whereas abundances in macrobial datasets were reported as counts of individuals, taxonomic abundance and identification of microbes in natural environments is commonly derived from DNA harvested from environmental samples; individual counts are not practical. Third, among macrobial datasets there are large differences in how communities were sampled (e.g. plot counts of trees, transects for breeding birds, multiple trapping/sampling methods for mammals) (see Appendix C).

Form of SADs in the Feasible Set

The canonical, hollow-curve, form of the species abundance distribution includes large numbers of rare species and small numbers of abundant species, leading to frequency distributions with the mode at small values of N and long, right-skewed, tails. To determine if this form is common in the set of possible SADs we analyzed the distribution of modal abundance class, species evenness, and skewness within the feasible set for a variety of N and S combinations and N/S ratios. We avoided extremely large values of S because values of S close to N are uncommon in nature and constrain the SAD to a nearly even vector of singletons. We used uniform random samples of 500 macrostates for each N-S combination in this analysis. These numbers are large enough to characterize the general form of the feasible set and small enough to permit doing so in

reasonable time (Fig. 1 of Appendix C).

Comparing Observed Data to Central Tendencies of Feasible Sets

We determined which SAD represented the center of each feasible set by generating 300 to 500 random macrostates from the feasible set (generating 500 random macrostates for some combinations of N and S was untenable). Random samples of 300, 500, and 700 macrostates produce equivalent results (Fig 1. of Appendix C). We chose the macrostate that overlapped the most on average with other random macrostates across the S ranked abundances. In the case of a tie, we favored the macrostate having the more evenly distributed overlap across ranked abundances (i.e. the macrostate with the smaller variance in overlap with other macrostates). This yielded SADs that were centered within the densest regions of random samples (Fig. 2 of Appendix C), and hence, within the central tendency of the feasible set. We compared this central SAD for each community to the observed SAD using rank-abundance distributions (RADs). Specifically we compared the observed value of abundance at each rank (most abundant to least abundant) at each site to the abundance at that same rank from the SAD representing the central tendency of the feasible set. We used log-transformed values of abundance at each rank (not log-transformed bins; see Nekola et al. 2008) to make visual comparisons and calculate R^2 values following Marks & Muller-Landau (2007) to avoid overweighting rare species, to address heteroscedasticity, and because we are generally more interested in proportional differences in abundance within a rank rather than absolute differences.

Results

The majority of possible SAD shapes exhibit the classic hollow-curve form with modes at low abundance classes and positive skewness, revealing an overall hollow-curve shape for most of the macrostates in the feasible set (Fig. 2—1, see also Fig 3 of Appendix C). The specific form of the distribution is influenced by the values of N, S, and average species abundance (i.e. N/S), which are associated with modal abundance, species evenness, and skewness of the SAD (Fig 2—2). This means that differences in community structure among sites (or directional changes along gradients) could result from the constraining influence of N and S. For realistic values of average abundance, the portion of highly uneven macrostates in the feasible set will increase as N is partitioned across a greater number of species. However, as average abundance approaches 1.0, the SAD must necessarily become highly even.

Observed ranked abundances were often similar to those near the central tendency of the feasible set, both within and across sites for trees, animals, and microorganisms (Fig 2—3). The SAD at the central tendency of the feasible set consistently explained the majority of variation in observed abundance distributions both within sites and among entire datasets (R^2: BBS = 0.93; CBC = 0.77; FIA = 0.84; GENTRY = 0.81; MCDB = 0.78; TERA = 0.83; AQUA = 0.58; FUNGI = 0.76; R^2 values are with respect to the central tendency, not a fitted relationship). However, clear deviations from the form of the central tendency did occur and were strongest among microbial metagenomes where the central tendency of the feasible set contained lower abundances for dominant species and higher abundances for rare species than the observed communities. We observed the

same pattern for microbes regardless of whether species were defined at 95, 97, or 99% (Fig. 4 of Appendix C).

Because many of the possible SADs are similar, the similarity between the center of the feasible set and the observed data means that the shapes of observed SADs tend to look very similar to the majority of possible shapes, suggesting a strong influence of the limits of observable variation on natural variation. Evaluating the correlation between the observed SAD and all random macrostates shows that randomly choosing a macrostate will often produce a distribution that is well correlated with observed data (Fig 4). This was most obvious for BBS, where the majority of randomly sampled macrostates explained more than 80% of observed variation in abundance for nearly all sites.

Discussion

The hollow-curve SAD has been referred to as an ecological law and is thought to be universal across taxa (McGill et al. 2007). This pattern is also observed in non-biological systems (Gaston et al. 1993; Nekola & Brown 2007; Warren et al. 2011) suggesting that the unevenness and ubiquity of the hollow-curve SAD might be explained by emergent statistical phenomena rather than specific biological processes (Šizling et al. 2009b; McGill 2010; White et al. 2012, Yen et al. 2012). Here, we have described the first attempt to understand the shape of the SAD in terms of the set of all possible shapes given two general constraints that are commonly used as inputs in ecological theory. The majority of feasible SADs share similar forms that, like observed SADs, resemble a hollow curve frequency distribution. As such, the feasible set provides an *a priori* reason for the ubiquity of the hollow-curve and a reason why many different models tend to

produce the same general SAD form. Examination of over 9,000 communities shows that observed SADs are often similar to the central tendency of the feasible set and, because most macrostates are clustered near the central tendency, the majority of possible distributions often explain substantial portions of variation in observed abundances.

While much of the variation in empirical SADs is characterized by the center of the feasible set, SADs are often more uneven than the central tendency. SADs for microbial communities were almost always exceptionally uneven, regardless of whether species were delineated at 95, 97, or 99% sequence similarity. Though hollow-curve SADs have been widely documented for microbes and macrobes, our examination reveals that the structure of microbial communities, with respect to the influence of N and S, may differ from that of macrobes. Indeed, microbial communities are known for their large rare portions (i.e. rare biosphere). However, it has also been suggested that the exceptional unevenness of microbial SADs may result from detection issues related to metagenomic methods that can exaggerate dominance and rarity (Woodcock et al. 2006). Observational/sampling biases are also a potential issue for the macrobial datasets (e.g., MacKenzie & Kendall 2002) and therefore have the potential to play a role in deviations from the feasible set in those analyses as well.

Empirical data can also be compared to the feasible set by comparing distributions of a statistical property (e.g. species evenness) across the feasible set. This allows the values for individual communities to be placed within context. For example, a community with a value of species evenness in the 50^{th} percentile of the feasible set, i.e. near the central tendency and the majority of possible macrostates, would not have an

exceptionally even or uneven distribution of abundance, regardless of whether the value of evenness itself is large or small (Fig 5 of Appendix C). This is particularly important when comparing sites that differ in N and S, since differences in evenness can be expected based purely on differences in the feasible set (Figure 2—2, Fig 5 of Appendix C). Consequently, comparisons of species evenness that do not account for the feasible set are primarily comparisons of N and S.

In addition to contextualizing single communities, ensembles of sites with shared values of N and S can be used to compare distributions of a property across communities to the distribution of that property in the feasible set. Conducting this analysis using FIA data and species evenness (E_{var}; Smith & Wilson 1996) reveals that, while the modal values of E_{var} for feasible sets and FIA sites were often similar, the distribution of E_{var} across the feasible set was broader than that of empirical distributions (Fig. 2—5).This relatively low natural variability could indicate that interactions between ecological processes and statistical phenomena prevent the extreme values of evenness that are otherwise possible. Additionally, the tendency for the distribution of empirical E_{var} values to be concentrated at lower or higher values of E_{var} was related to average abundance (i.e. N/S), with higher N/S leading to lower E_{var} (i.e. lower evenness) for both empirical SADs and the feasible set. While the general decrease in species evenness with average abundance can be explained by the feasible set (Fig. 2—2), the actual change in empirical E_{var} outpaced that of the feasible set (Fig. 2—5), suggesting that mechanisms leading to unevenness may strengthen as N/S increases (e.g. via positive frequency dependence), but not so much that the lowest possible range of species evenness is attained.

While the feasible set reveals that a small number of community-related constraints may explain the general shape of the SAD by limiting observable variation, it also demonstrates that in some cases empirical patterns deviate directionally from the majority of possible states (Fig. 2—3) and are more tightly clustered than expected (Fig. 2—5). Consequently, the ecological interactions of individuals, populations, and species may be needed to explain the specific form of ecological patterns as well as the frequent occurrence of exceptionally uneven SADs and the rare occurrence of exceptionally even ones. High degrees of competition and dispersal limitation, and low degrees of invasiveness may all lead to the degrees of excessive dominance or unevenness that are commonly observed among microorganisms and macroorganisms and which cannot be attributed to the constraining influences of N and S. However, without the feasible set it would not possible to recognize that this degree of unevenness and its relatively low natural variability are exceptional.

The feasible set approach focuses on the observable variation among the possible forms of a pattern of interest (i.e. macrostates). In a way, it assumes that all possible forms of the SAD are equally likely because it assumes nothing about the ways in which each macrostate may arise. However, by accounting for the ways in which macrostates can arise through microstate configurations, approaches like MaxEnt (Pueyo et al. 2007; Harte et al. 2008; Frank 2011) produce a most likely form that may better explain the general shape of the SAD. Indeed, 4 out of the 5 datasets shared with White et al. (2012) are at least somewhat better fit by the predictions of the MaxEnt model of Harte (2011); the exception being BBS. This comparison is approximate because we worked with

subsets of the datasets in White et al. (2012) and because the model of Harte (2011) requires additional assumptions to be made beyond fixing N and S.

The idea that empirical SADs may be more similar to the form with the greatest number of microstates than to the form closest to the center of the feasible set is complicated by the fact that MaxEnt yields different predictions depending on the specific approach to the problem (Haegeman & Etienne 2010). In cases where the number of constraints is small, it is unlikely that the most likely macrostate from one of the several MaxEnt approaches will occur at the center of the feasible set. In fact, Haegeman & Loreau (2008) consider differences between MaxEnt predictions and the center of the feasible set to be a necessary condition for applications of MaxEnt to be considered non-trivial. This presents an interesting philosophical question: should we try to understand patterns in the context of their distribution of macrostates alone, in the context of these macrostates weighted by the number of microstates, or some combination of the two. Current microstate-based approaches do not explicitly consider the properties of either the feasible set or the full set of microstates, only a single most likely macrostate. This prevents existing MaxEnt approaches from providing a general context for how extreme an abundance distribution is relative to the most likely macrostate, though this can probably be addressed through sampling approaches to randomly select microstates. Further research is needed to compare and understand the relationships between these microstate and macrostate based-approaches, to form a more comprehensive understanding of how to contextualize empirical patterns and theoretical predictions.

Another area for additional research is understanding what functional forms (e.g. log-series, log-normal, etc.; McGill et al. 2007) are most common in the feasible set and whether the most common forms change as a function of S and N. This would provide information useful for comparing the quality of distributional fits to empirical data. It would also provide context for one of the current challenges for theoretical models of macroecological patterns - making predictions that are valid at multiple taxonomic and spatial scales (e.g, Šizling et al. 2009a). In contrast to most theoretical models, it is possible that the form of the central tendency changes with N, S, and N/S, becoming more or less similar to different standard distributions (e.g. log-series, log-normal). Knowing how the feasible set responds to changes in N (e.g. with sample size and area) and S (e.g. different taxonomic levels) could enlighten the discussion of whether universal forms of macroecological patterns exist (Šizling et al. 2009a). It has been suggested that, as N approaches infinity and as S changes as a function of N (i.e. $S = cN^{1/2}$), there is a limit shape to random integer partitions (Vershik & Yakubovich 2001). Further studies are needed to explore whether this is the case for combinations of N and S observed in natural systems, and whether this limit shape is similar to known distributions.

The feasible set approach is part of an emerging area of ecology that uses constraint or state-variable-based approaches to understand ecological patterns (Shipley et al. 2006; Pueyo et al. 2007; Haegman & Loreau 2008; McGill 2010; Harte 2011). These approaches take a top down perspective on understanding ecological patterns, suggesting that much of the information contained in a distribution can be captured by a

small number of constraints. This approach to understanding ecological patterns has received empirical support from observational (Harte 2011; White et al. 2012) and experimental (Supp et al. 2012) studies. However, even when these approaches successfully characterize empirical patterns, they do not indicate whether ecological processes are not operating. Instead ecological processes may influence emergent patterns indirectly through their influence on constraints or state variables (White et al. 2012; Supp et al. 2012). These constraint-based approaches reinforce the fact that ecological processes operate within but also influence constraints that necessarily determine a set of possible outcomes.

The feasible set represents a new perspective in understanding empirical patterns. This approach is potentially applicable to many other widely-known distributions in ecology and other areas of science. In particular, the SAD is a specific type of distribution of wealth and uneven distributions of wealth are widespread in social, economic, and physical systems (Zipf 1949, Gaston et al. 1993, Reed 2001, Nekola & Brown 2007). The feasible set approach should be applicable to distributions of wealth and abundance that are characterized by the partitioning of a total quantity (e.g. individuals, species, dollars, hectares) among a number of classes (e.g. species, islands, socioeconomic classes, countries). This includes classic ecological patterns, such as the species-area relationship and species-time relationship, and emerging patterns in microbial ecology such as distribution of functional traits, as well as distributions of wealth, size, and abundance among human populations

References

Amend, A.S., Seifert, K.A., Samson, R. & Bruns, T.D. (2010). Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *P. Natl. Acad. Sci. USA.*, 107, 13748-13753.

Andrews, G.E. & Eriksson, K. (2004). *Integer Partitions*. Cambridge Univ. Press, New York.

Chu, H., Fierer, N., Lauber, C.L., Caporaso, J.G., Knight, R. & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol*., 12,2998–3006.

Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A. & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.*, 6, 1007–17.

Flores, G.E., Campbell, J., Kirshtein, J., Meneghin, J., Podar, M., Steinberg, J.I., *et al.* (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol.*, 13, 2158-2171.

Frank, S.A. (2011). Measurement scale in maximum entropy models of species abundance. *J. Evolution. Biol.*, 24, 485-496.

Gaston, K.J., Blackburn, T.M. & Lawton, J.H. (1993). Comparing animals and automobiles: a vehicle for understanding body size and abundance relationships in species assemblages? *Oikos*, 66, 172-179.

Haegeman, B. & Etienne, R.S. (2010). Entropy maximization and the spatial distribution of Species. *Am. Nat.*, 175, E74-E90.

Haegeman, B. & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, 117, 1700-1710.

Harte, J. (2011). *Maximum Entropy and Ecology*. Oxford Univ. Press, Oxford.

Harte, J., Zillio, T., Conlisk, E., & Smith, A.B. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology*, 89, 2700-2711.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton Univ. Press, Princeton.

Lamendella, R., Domingo, J.W.S., Ghosh, S., Martinson, J. & Oerther, D.B. (2011). Comparative fecal metagenomics unveils unique functional capacity of the swine gut. *BMC Microbiol.*, 11, 103.

Lazarevic, V., Whiteson, K., Hernendez, D., François, P. & Schrenzel, J. (2009). Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics*, 11, 523.

MacKenzie, D. I. & Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology, 83*, 2387-2393.

Marks, C.O. & Muller-Landau, H.C. (2007). Comment on "from plant traits to plant communities: A statistical mechanistic approach to biodiversity". *Science*, 316, 5830.

Marshal, M.M., Amos, R.N., Henrich, V.C. & Rublee, P.A. (2008). Developing SSU rDNA metagenomic profiles of aquatic microbial communities for environmental assessments. *Ecol. Indic.*, 8, 442-453.

Martiny, J.B.H, Eisen, J.A., Penn, K., Allison, S.D. & Horner-Devine M.C. (2011). Drivers of bacterial β-diversiy depend on spatial scale. *P. Natl. Acad. Sci. USA*, 108, 7850-7854.

McGill, B.J. (2010). Towards a unification of unified theories of biodiversity. *Ecol. Lett.*, 13, 627–642.

McGill, B.J. & Collins, C. (2003). A unified theory for macroecology based on spatial patterns of abundance. *Evol. Ecol. Res.*, 5, 469-492.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., *et al.* (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, 995–1015.

McGill, B.J. & Nekola, J.C. (2010). Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos*, 119, 591–603.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al.* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386.

National Audubon Society. (2002). The Christmas Bird Count historical results. Retrieved from http://www.audubon.org/bird/cbc.

Nekola, J.C. & Brown, J.H. (2007). The wealth of species: ecological communities, complex systems and the legacy of Frank Preston. *Ecol. Lett.*, 10, 188-196.

Nekola, J.C., Šizling, A.L., Boyer, A.G. & Storch, D. (2008). Artifactions in the log-transformation of species abundance distributions. *Folia Geobot*., 43, 259-268.

Nijenhuis, A. & Wilf, H.S. (1978). *Combinatorial Algorithms for Computers and Calculators*. Academic Press, New York.

Pielou, E. 1975. *Ecological Diversity*. Wiley, New York.

Pueyo, S., He, F. & Zillio, T. (2007). The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol. Lett.*, 10, 1017-1028.

Reed, W. J. 2001. The Pareto, Zipf, and other power laws. Economic Letters 74:15-19.

Roselló-Morak, R. & Amann R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.*, 2001, 39-67.

Sauer, J.R., Hines, J.E., Fallon, J.E., Parkieck, D.J., Ziolkowski, D.J. Jr. & Link, W.A. (2011). *The North American Breeding Bird Survey 1966-2009*. Version 3.23.2011. USGS Patuxent Wildlife Research Center, Laurel, MD.

Phillips, O. & Miller, J.S. (2002). *Global Patterns of Plant Diversity: Alwyn H. Gentry's Forest Transect Data Set*. Missouri Botanical Garden Press, St. Louis, Missouri, USA.

Schloss, P.D. & Handelsman, J. (2006). Toward a census of bacteria in soil. *PLoS Comput. Biol.*, 2, e92.

Shipley, B., Vile, D. & Garner, É. (2006). From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science*, 314, 812-814.

Šizling, A.L., Storch, D., Reif, J. & Gaston, K.J. (2009a). Invariance in species-abundance distributions. *Theor. Ecol*., 2009, 89-103.

Šizling, A.L., Storch, D., Šizlingova, E., Reif, J. & Gaston, K.J. (2009b). Species abundance distribution results from a spatial analogy of central limit theorem. *P. Natl. Acad. Sci. USA*.,106, 6691-6695.

Smith, B. & Wilson, J.B. (1996). A consumer's guide to evenness indices. *Oikos*, 76, 70-82.

Sugihara, G. (1980). Minimal community structure: an explanation of species abundance patterns. *Am. Nat*., 116, 770–787.

Supp, S.R., Xiao, X., Ernest, S.K.M. & White, E.P. (2012). An experimental test of the response of macroecological patterns to altered species interactions. *Ecology*, 93, 2505-2511.

Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and abundance of mammalian communities. *Ecology*, 92, 2316-2316.

Ulrich, W., Ollik, M. & Ugland, K.I. (2010). A meta-analysis of species–abundance distributions. *Oikos*, 119, 1149–1155.

U.S. Department of Agriculture, F.S. (2010). Forest inventory and analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis.

Vershik, A. & Yakubovich, Y. (2001). The limit shape and fluctuations of random partitions of naturals with fixed number of summands. *Mosc. Math. J.*, 1, 457-468.

Warren, R.J. II., Skelly, D.K., Schmitz, O.J. & Bradford, M.A. (2011). Universal Ecological Patterns in College Basketball Communities. *PLoS ONE*, 6, e17342.

Webster, N.S., Taylor, M.W., Behnam, F., Lücker, S., Rattel, T., Whalan, S., *et al*., (2010). Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environ. Microbiol*., 12, 2070-2082.

White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

Woodcock, S., Curtis, T.P., Head, I.M., Lunn, M. & Sloan, W.T. (2006). Taxa–area relationships for microbes: the unsampled and the unseen. *Ecol. Lett.*, 9, 805–812.

Yen, J.D.L, Thompson, J.R. & MacNally, R. (2012). Is there an ecological basis for species abundance distributions? *Oecologia*, 10.1007/s00442-012-2438-1.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Oxford.

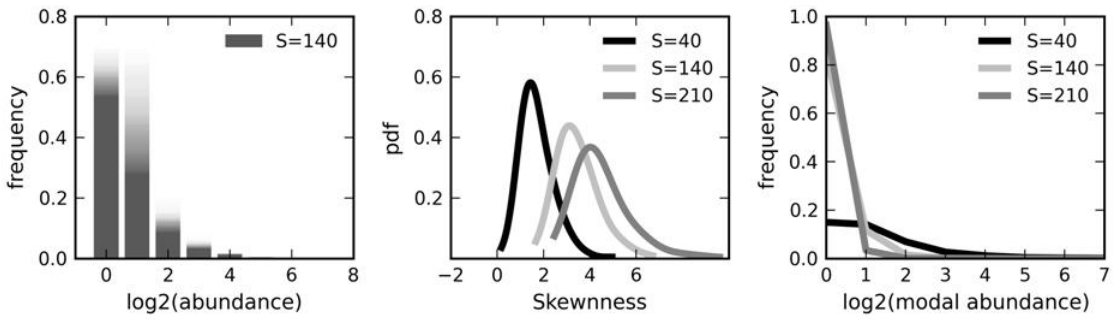


Figure 2 – 1. Plots revealing characteristics of the species abundance distribution feasible set. Left, plots of 500 randomly sampled macrostates in the feasible set for N = 1,000 and S = 140. Each macrostate is plotted as a light grey frequency distribution of $\log_2$ abundance classes. Overlap of these distributions produces the gradual shading to dark grey. Center and Right, plots of skewness and modal abundance across random samples of feasible set for N = 1000 and S = 40, 140, 210 reveal that feasible sets are dominated by right-skewed macrostates and that the modal abundance class tends towards singletons or small abundances, indicating that feasible sets are dominated by similarly-shaped hollow-curves.

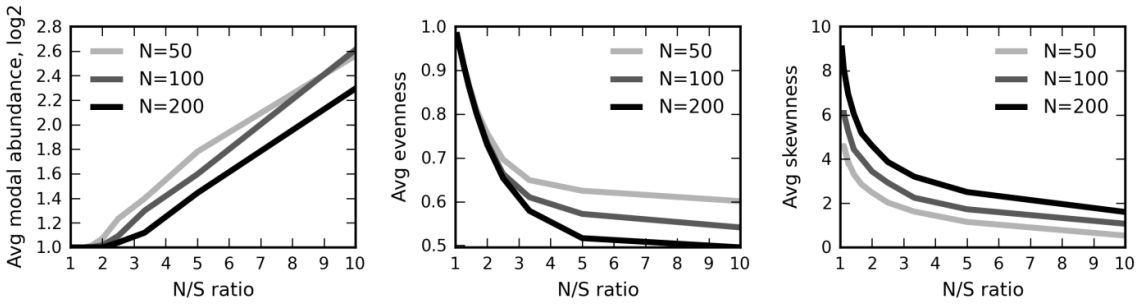


Figure 2 – 2. Plots revealing how feasible set characteristics change with average abundance. Plots of average abundance (N/S) against modal abundance, evenness, and skewness averaged across 500 randomly sampled macrostates for N = 50, 100, 200 and S = {N/10, N/9, …, N}. The monotonic change in these features of the feasible set with increasing N/S across doublings of N suggests predictable changes and constraints on community structure resulting from changes in N and N/S.

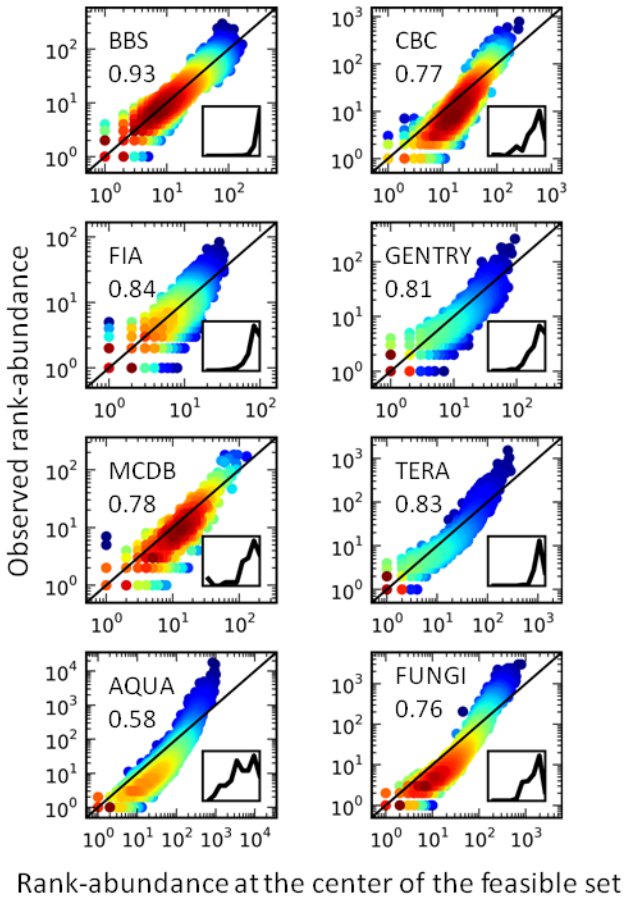


Figure 2 – 3. Plots of the relationship between observed rank-abundances from all sites in a dataset and the corresponding ranked abundances at the center of the feasible set. Each point represents a rank in a community with the y-coordinate showing the observed abundance at that rank and the x-coordinate showing the abundance at the center of the feasible set. Data are heat mapped to reveal the density of rank-abundance states, which is largely centered around the 1:1 line for some datasets (e.g. BBS, GENTRY) and deviates more greatly for others (e.g. AQUA, FUNGI). Insets are of kernel density curves for site specific R^2 values; the x-axis ranges from 0.0 to 1.0.

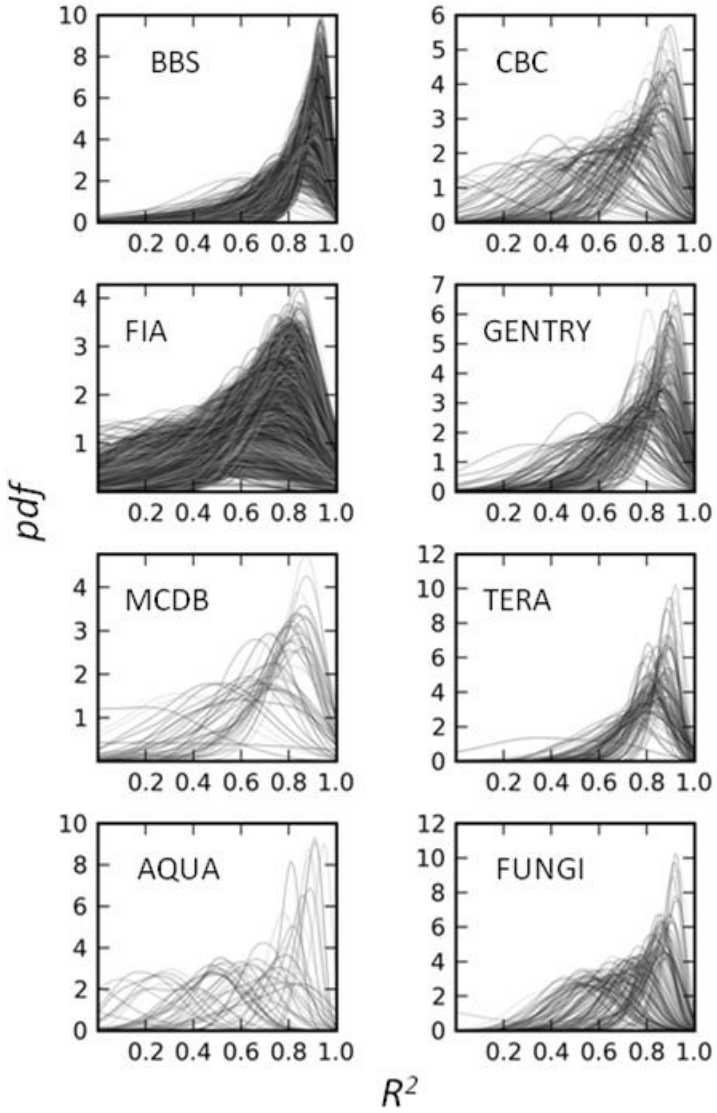


Figure 2 – 4. Kernel density curves of R^2 values relating random macrostates to the observed RAD as in Figure 3. Each site is represented by a single kernel density curve, revealing that the majority of a random sample of the feasible set often describes large portions of variation in ranked abundances at a site.

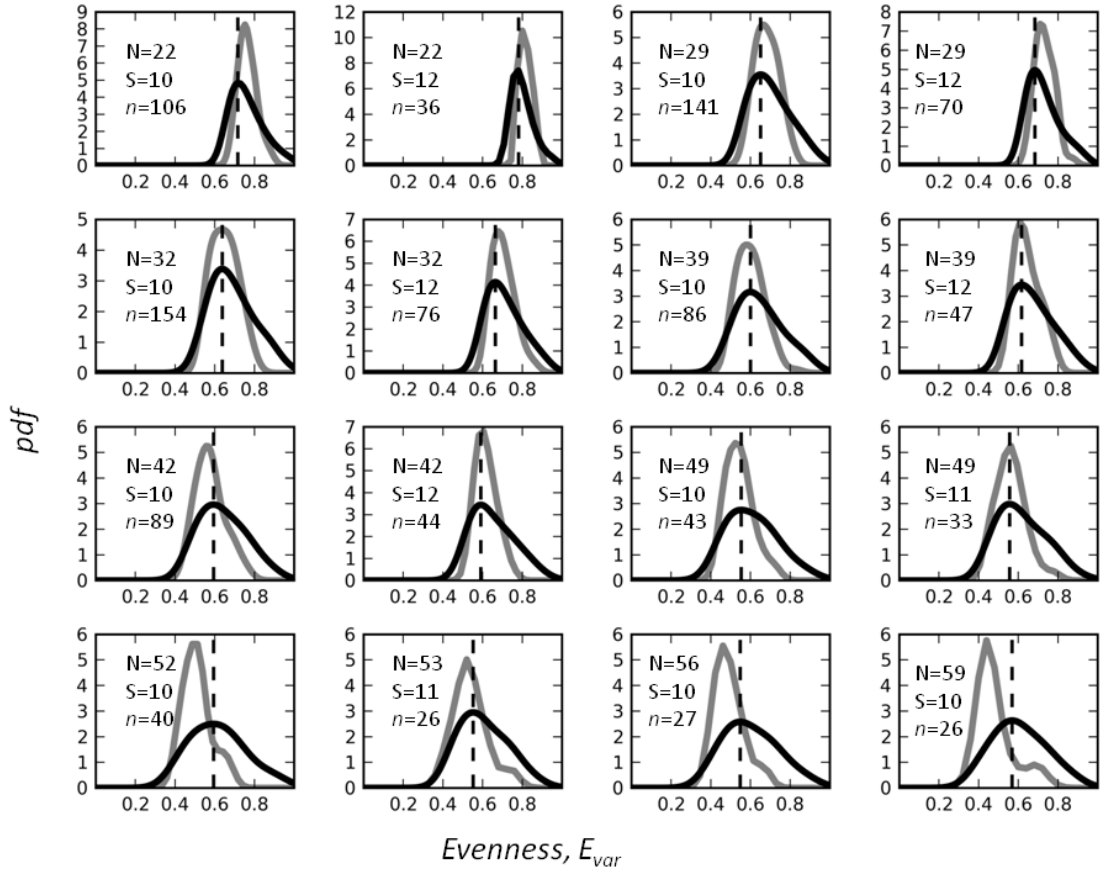


Figure 2 – 5. Plots of kernel density curves for E_{var} across entire feasible sets (black curves) and kernel density curves for E_{var} across sites in FIA with the same N and S (grey lines). Sample size, i.e. number of FIA sites, is given as 'n'. Modes of the feasible sets are shown by vertical dashed lines. Each column reveals 1.) a shift in the mode of the feasible set towards lower evenness as average abundance (i.e. N/S) increases as in Figure 2; 2.) a shift in the distribution of empirical E_{var} towards lower evenness that out-paces the changing mode of the feasible set; and 3) a more narrow distribution of observed E_{var} values than expected from sampling from the feasible set.

CHAPTER 3

EFFICIENT ALGORITHMS FOR SAMPLING FEASIBLE SETS OF MACROECOLOGICAL PATTERNS

Abstract

Ecological variables such as species richness (*S*) and total abundance (*N*) can strongly influence the forms of macroecological patterns. For example, the majority of variation in the species abundance distribution (SAD) can often be explained by the majority of possible forms having the same *N* and *S*, i.e. the feasible set. The feasible set reveals how variables such as *N* and *S* determine observable variation and whether empirical patterns are exceptional to the majority of possible forms. However, this approach has currently only been applied to the SAD using relatively inefficient random sampling algorithms. We extend the use of the feasible set approach by developing new algorithms to efficiently generate random samples of the feasible set for the SAD and the intraspecific spatial abundance distribution (SSAD). These algorithms are often several orders of magnitude faster than a previous method, which greatly increases the size and diversity of communities that can be examined.

*Coauthored by: Locey, K.J. & McGlinn, D.J. (**2013**) Efficient algorithms for sampling feasible sets of macroecological patterns. *PeerJ PrePrints*, 1, e78v1 http://dx.doi.org/10.7287/peerj.preprints.78v1

Introduction

Understanding patterns of abundance, distribution, and diversity is a primary goal in ecology (Brown 1995). These macroecological patterns reveal general characteristics of ecological systems and provide insights into the processes and phenomena that drive ecological structure. It has also been shown that general ecological variables such as total community abundance (*N*) and species richness (*S*) can greatly constrain the forms of ecological patterns (Harte et al. 2008, McGlinn & Hurlbert 2012, Supp et al. 2012, White et al. 2012, see also chapter 2). In fact, more than 90% of observed variation in the species abundance distribution (SAD), i.e. the vector of abundances of all species in a community (McGill et al. 2007), can often be explained by models constrained by *N* and *S* (Harte 2011, White et al. 2012). This is perhaps not surprising given that the majority of all possible forms of the SAD having the same *N* and *S*, i.e. the feasible set, often explain the general form of empirical SADs (see also chapter 2).

The feasible set reveals how constraint variables such as *N* and *S* determine observable variation and whether empirical patterns are exceptional to or representative of the majority of possible forms having the same constraint values (Haegeman & Loreau 2008, see also chapter 2). Currently only the SAD has been examined using the feasible set approach (but see Storch et al. 2008). Though results from chapter 2 suggested that other macroecological patterns could also be examined in the context of their feasible sets, there are considerable computational challenges to using a feasible set approach.

Feasible sets can be immense and enumerating them can be untenable. However, small random samples can be used to characterize the center of the feasible set (i.e.

average form) as well as the distribution of statistical features (e.g. evenness, diversity) within the feasible set. In chapter 2 we took a conceptually simple and unbiased approach to sampling the SAD feasible set, a combinatorial method known as integer partitioning. This approach is based on the fact that there are a limited number of unordered ways that n integers can sum to a total q, and hence, a limited number of ways that the abundances of S unlabeled species can sum to a total abundance of N. These unordered configurations of integers are called integer partitions (Bóna 2006). For example, the feasible set for $q = 6$ and $n = 3$ is: {(4, 1, 1), (3, 2, 1), (2, 2, 2)}, where differently ordered configurations having the same integer values (e.g. (4, 1, 1), (1, 1, 4), (1, 4, 1)) represent the same integer partition, i.e. (4, 1, 1).

Use of integer partitioning to randomly sample the feasible set allows the feasible space to be characterized without generating all possible forms. However, all published partitioning algorithms sample the feasible set only with regards to the total, q. This means that all partitions of q, regardless of the number of elements n, have the same probability of being drawn. For example, randomly generating one partition for $q = 1000$ and $n = 10$ requires drawing from a feasible set of nearly 2.4×10^{31} partitions, one of the roughly 8.9×10^{14} having 10 elements; a probability of nearly 3.7×10^{-17}. This means that randomly sampling the feasible set for a given q and n, requires generating partitions according to q and then rejecting those not having n elements, often resulting in impractically high rejection rates.

Another challenge in applying integer partitioning to the study of feasible sets is that some macroecological patterns include parts with zero values. One example is the

species spatial abundance distribution (SSAD) describing the frequency with which individuals of a single species occupy areas within a landscape (Brown et al. 1995, Harte et al. 2008). In the SSAD, individuals can be absent from a number of areas, meaning that there are some areas with zero individuals. Because integer partitions *per se* do not include zeros, integer partitioning methods need to be modified to examine the SSAD feasible set.

Here, we present algorithms that greatly increase the efficiency of sampling the SAD and SSAD feasible sets. We explain each algorithm in concept and develop Python and R based implementations of them. We test each algorithm for sampling bias and for speed against the method of chapter 2. To reveal the practical gains of these new algorithms, we reanalyze the SAD datasets of chapter 2 wherein it took more than 10000 compute hours to examine only 60% of the available data (9562 of 15950 SADs). Our algorithms should allow us to examine a much larger portion of the data in less time. Finally, we examine general characteristics of the SSAD feasible set including characteristics of the frequency distribution, skewness of the SSAD, and the rank distribution of abundances. Our work expands the feasible set approach to an additional ecological pattern and to values of q and n that were previously untenable.

Methods

Our goal is to develop fast and unbiased integer partitioning algorithms, to generate random samples of feasible sets for ecological patterns such as the SAD and SSAD that are defined by a total q and composed of n elements. Integer partitioning is a

mature field of mathematics and algorithms for generating random partitions of q (e.g. Nijenhuis and Wilf 1978) are often implemented in mathematical environments (e.g. Sage, Mathematica). However, these approaches do not allow the random partitioning of q into exactly n elements. Here, we use two well-established theorems and a partitioning identity to develop a general method for generating a random partition of q having n elements. We then modify our method to allow elements having zero values. We begin by using two theorems to generate a random partition of q (see Appendix D for generating functions and recurrence relations).

1) For every integer q there are $p(q)$ partitions having q or less elements.
2) For every integer q there are $p_k(q)$ partitions having k or less as the first element.

For example, there are 5 partitions of 4 and each has 4 or less as the first element; i.e. $p(4) = 5$, {(4), (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)}. Likewise, there are 4 partitions of 4 having 3 or less as the first element; $p_3(4) = 4$, {(3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)}. It is possible to sequentially build a random partition element by element by iteratively applying these two theorems (Fig. 3—1). Specifically, if we choose a random number x from 1 to $p(q)$, we can say there are at least x partitions of q having some value k or less as the first element, i.e. $x \leq p_k(q)$. Likewise, there must also be some value of $k - 1$, for which, there are less than x partitions of q having $k - 1$ or less as the first element, i.e. $p_{k-1}(q) < x$. Putting these statements together, there must be a value k for which $p_{k-1}(q) < x \leq p_k(q)$. In this way, we can find the value of the first element in one of the partitions by finding the value of k that satisfies $p_{k-1}(q) < x \leq p_k(q)$. Having found the

value of the first element, we can decrease x by $p_{k-1}(q)$ and q by k, and then find the first element for this combination of smaller values. Repeating this process will sequentially build the partition until $q = 0$ and the sum of the partition is equal to the original value of q (Fig 3—1).

The above approach is similar to well-established methods of generating random partitions of q (e.g. Nijenhuis & Wilf 1978, Stojmenovic 2008). However, our goal is to generate random partitions of q having n elements. For this, we use a well-known integer partitioning identity to restrict the number of elements in a randomly chosen partition to n. Specifically, the number of partitions of q having n elements equals the number of partitions of q having n as the first element (Bóna 2006). This is because each partition of q having n elements corresponds to one unique partition of q having n as the first element (Bóna 2006). For example, consider the partition (3, 1), which can be illustrated with rows of dots, called a Ferrer's diagram (Fig 1). In the Ferrer's diagram for (3, 1) there are two rows, the largest having three dots. Flipping the diagram on its diagonal produces its conjugate (2, 1, 1), which has three rows, the largest row having two dots. So, the conjugate of (3, 1) is (2, 1, 1) and vice versa. Consequently, the first part of an integer partition determines the number of parts in its conjugate (Bóna 2006). This allows us to extend the problem of generating random partitions of q to random partitions of q having exactly n elements. That is, knowing the first element must be n so that its conjugate has n elements, we can decrease q by n and then generate a random partition for this decreased value of q having n or less as the first element. Once the partition is generated, we append n to the beginning of the

partition and conjugate it to produce a random partition of the original q having exactly n elements (Fig 3—1).

The approach outlined above begins with a randomly chosen number x between 1 and $p(q)$.. It then finds the value of k that satisfies the inequality $p_{k-1}(q) < x \leq p_k(q)$, that value of k is the value of the first element of the partition. However, the question remains as to which value of k to start with and how to proceed to different values. Indeed, we could start with the smallest possible value of k ($k = 1$) and take a 'bottom-up' approach, or the largest possible value ($k = p(q)$) and take a 'top-down' approach, or even choose k at random and use a 'divide-and-conquer' method. These approaches differ only in how k is chosen and each builds the partition one element at a time. However, in the event that q is much larger than n, e.g. all trees in the 50 ha Barro-Colorado Island mapped forest plot where $q \approx 200{,}000$ individuals and $n \approx 300$ species, the three algorithms above will still be inefficient. This is because they would first generate a partition of, say, 200000 having 300 as the first element, but having as many as 199701 elements, and then conjugate it to produce a partition of 200000 having 300 elements. Clearly building the partition one element at a time would be inefficient in this case.

An alternative approach is to build a partition using multiples of integers – the 'multiplicity' approach. Instead of finding the value of k corresponding to the first element, appending it and moving on, we can instead ask how many times must k occur, i.e. the partitions having some multiple m of k. We can start with the smallest possible multiple (i.e. $m = 1$) and ask whether x is less than or equal to the number of partitions of $q - k*m$ having less than k as the first part. This is because the set of partitions of q

having a number of k's equal to m actually contains the set of partitions of $q - k*m$ having less than k as the first part (Appendix D). We can increase m by one until $x \leq p_k(q - k*m)$, at which point we will have found the corresponding multiple of k. One drawback to this method is the overhead due to the extra computation.

Random partitions for q and n, with some parts having zero values

The above algorithms address patterns having positive values, e.g. SAD. In contrast, some macroecological patterns include zero values (e.g. absences). One example is the species spatial abundance distribution (SSAD), a frequency distribution that characterizes the number of quadrats, cells, or areas containing a given abundance of a species (Brown et al. 1995). However, only small changes are needed to adapt the above approaches to cases allowing zero-valued parts. For example, let 10 unlabeled individuals occupy a landscape sectioned into quarters. The most aggregated distribution would be for all 10 to occupy the same quarter, [10, 0, 0, 0]. The least aggregated would be for 3 to occupy two quarters while 2 occupy the other two quarters, [3, 3, 2, 2]. In fact, the number of configurations for 10 unlabeled individuals distributed across 4 unlabeled sections equals the number of partitions of 10 having 4 or less parts, i.e. $p_4(10 + 4) = 23$. Consequently, if $n \leq q$, a random partition for q and n allowing for zero-valued parts, is simply a random partition for q having n or less parts, with zeros appended to ensure the final form of the partition has n parts.

On the other hand if $n > q$ a different approach is needed. To see this let 4 unlabeled individuals occupy a landscape sectioned into tenths. The most aggregated distribution would be for all 4 to occupy the same subsection, [4, 0, 0, 0, 0, 0, 0, 0, 0, 0] and the least

aggregated configuration would be for 4 sections to have one individual and for 6 sections to have zero, i.e. [1, 1, 1, 1, 0, 0, 0, 0, 0, 0]. In this way, the number of possible configurations for 4 unlabeled individuals distributed across 10 unlabeled sections is $p(q)$. Consequently, if $q < n$, a random partition for q and n, allowing for zero-valued parts, is simply a random partition for q having q or less parts, with zeros appended to ensure the partition has n parts.

Examining for bias and speed

We implemented the above algorithms in Python and R and made them freely available using a public Github repository (https://github.com/klocey/partitions). We are currently developing these algorithms into a Python module to be distributed on the Python Package Index (PyPI; https://pypi.python.org/pypi) and an R package to be distributed on The Comprehensive R Archive Network (CRAN; http://cran.us.r-project.org/). We used kernel density curves to visually compare the results of the above algorithms to full feasible sets and to random samples generated with the function implemented in Sage, which is based on the algorithm of Nijenhuis & Wilf (1978) and is the method used in chapter 2. If our algorithms are unbiased, then their distributions will not differ in any systematic way from full feasible sets and random samples generated using Sage. We compare the computational speed of our algorithms to that of the approach used in chapter 2 (i.e. using Sage to generate random partitions for a given q and rejecting those not having n elements) across a range of values of q, n, and q-n ratios for which the latter method was likely to return random
samples within reasonable time (one hour).

Empirical Demonstration of the New Algorithms

In chapter 2 we analyzed 9562 SADs of trees, birds, mammals, fungi, and prokaryotes using a partitioning algorithm that sampled the feasible set according to total abundance *N* but not with respect to species richness *S* (i.e. the number of elements). Those data consisted, in part, of a subset of previously compiled datasets of site-specific species abundance data (see White et al. 2012), and included four continental-to-global scale surveys, including the Christmas Bird Count (129 sites) (National Audubon Society 2002), North American Breeding Bird Survey (1,586 sites) (Sauer et al. 2011), Gentry's Forest Transect Data Set (182 sites) (Phillips & Miller 2002), Forest Inventory Analysis (7,359 sites) (U.S. Department of Agriculture 2010), and one global-scale data compilation, the Mammal Community Database (42 sites) (Thibault et al. 2011). In chapter 2 we also compiled abundance data at the species level from five microbial metagenome projects for a total of 264 SADs. Those data were obtained from the metagenomics server MG-RAST (Meyer et al. 2008). Metagenomic data were compiled into datasets representing aquatic prokaryotic communities (48 metagenomes) (Flores et al. 2011, www.catlin.com/en/Responsibility/CatlinArcticSurvey), terrestrial prokaryotic communities (92 metagenomes) (Chu et al. 2010, Fierer et al. 2012), and terrestrial fungal communities (124 metagenomes) (Amend et al. 2010).

The inefficiency of the partitioning method used in chapter 2 restricted our analyses to combinations of *N* and species richness *S*, for which, there was a reasonable probability of generating a random integer partition of *q* with exactly *n* elements ($> 10^{-6}$). This restriction allowed for only 60% of the available data to be examined despite more

than 10000 compute hours worth of effort. We reanalyze those datasets using the algorithms developed here, which should allow for random samples of a greater number of SADs to be produced in less time.

General characteristics of the SSAD feasible set

Brown et al. (1995) revealed evidence that the general form of the SSAD, like that of the SAD, is characterized by a hollow-curve. In the sense of the SSAD, a hollow-curve implies that many areas are occupied by few or no individuals and that relatively few areas are occupied by many individuals. We generated random samples of the feasible set of the SSAD for ecologically realistic combinations of q and n, and examined their general features.

Results

Statistical properties of entire feasible sets are indistinguishable from random subsets generated with our sampling algorithms, demonstrating that the implementation of our algorithms was unbiased (Fig 3—2 and Figs 1 and 2 of Appendix D). When generating 300 random partitions, i.e. enough to safely characterize the feasible space (see chapter 2), these implementations were, at worst, one-to-two orders of magnitude faster than the method used in chapter 2 and were, at best, 4 to 5 orders of magnitude faster for the combinations of q and n we tested (Fig 3). These combinations were limited to values for which the algorithm used in Sage could generate random samples in reasonable time. Consequently, the algorithms we developed quickly produce random samples for values of q and n that are impractical with algorithms that sample only

according to *q*. Each algorithm was best suited for particular values of *q* and *n* (Fig 3—4). For cases where all parts have positive values, the multiplicity algorithm is the fastest for combinations where *q* is partitioned among a relatively small number of elements (Fig 3 of Appendix D).

The greater efficiency of the algorithms developed here allowed us to generate between 300 and 500 random partitions for 92.7% of the SADs (14786/15950) from the compilation of SAD data used in chapter 2, in less than 1000 compute hours. In contrast, the method used in chapter 2 required more than 10000 compute hours to generate between 300 and 500 random partitions for 60% of the available data (9562/15950 SADs).

Our examination of the SSAD feasible set supports the observation of Brown et al. (1995) that SSADs are characterized by hollow-curves (few cells with many individuals and many cells with few individuals) (Fig 5). The hollow-curve nature of the members of the feasible set increases as the total *q* is distributed across a greater number of elements (e.g. quadrats, areas).

Discussion

The feasible set approach is a general method for understanding how constraints determine observable variation in the forms of macroecological patterns and in distributions of wealth, in general (see chapter 2). We used integer partitioning to understand how the feasible set is ordered, to find the size and general features of the feasible set, and to generate unbiased random samples of the feasible set for a given a total (e.g. total community abundance, total species abundance) and number of entities

(e.g. species, quadrats). The algorithms we derived greatly increase the practical use of feasible set by decreasing computing time. In addition to examining the SAD, we expanded the feasible set approach to distributions with zero values, such as the SSAD. We also provided the algorithms in two computing languages frequently used by ecologists, R and Python, and have taken steps to ensure our implementations are unbiased.

Integer partitioning is only one way to examine and randomly sample the feasible set of possible SAD and SSAD shapes. Other possibilities include linear programming and iterative random walks, such as that used by Haegeman & Loreau (2008). Those approaches may not require combinatorial problems to be solved, i.e. $p_k(q)$, and so may not suffer from the problem of combinatorial explosion (large increases in the size of the feasible set for small changes in the total q and number of elements), as is the case with our algorithms. However, as stated in chapter 2 one benefit to the integer partitioning approach is that random sampling algorithms can be derived that are inherently unbiased and do not require 'burn-in' periods to produce effectively independent samples. The combinatorial approach also reveals properties such as the size of the feasible set, mathematical properties and symmetries within the feasible set (e.g. the theorems and identity used in this study), and a way to relate the feasible sets of different patterns, e.g. SAD and SSAD, in purely quantitative terms. However, we suggest that the combinatorial algorithms developed here should be compared to approaches such as those in Haegeman & Loreau (2008).

Our examination of the SSAD feasible set (Fig. 3—5) reveals that the central tendency of the set is characterized by a hollow-curve which is consistent with the empirical SSADs found by Brown et al. (1995). In that study, the authors state that the highly 'clumped' and hollow-curve nature of the SSAD resembles distributions used to predict the form of the SAD. The authors offer an ecological interpretation for the similarity between the patterns in terms of niche requirements. However, the SSAD feasible set (in purely mathematical terms and without interpreting what the total q and number of elements n represent) differs from the SAD feasible set only in that zero values are allowed, which also allows for $n > q$. Consequently, the two patterns are not only coupled by ecological mechanisms, but are also coupled by the purely mathematical properties of their feasible sets.

The feasible set approach ignores biological and statistical mechanisms and focuses entirely on observable variation in the shape of empirical patterns. Consistency of empirical patterns with the center of the feasible set suggests that the shapes of those patterns contain little information beyond that encoded by the constraints used to characterize the feasible set (Haegeman & Loreau 2008, see chapter 2). However, consistency with the feasible set does not mean that biological processes are not operating but rather that they may indirectly influence empirical patterns through their effects on constraints (Supp et al. 2012, White et al. 2012). Alternatively if empirical patterns occupy an uncommon portion of the feasible set (e.g. in being exceptionally uneven) biological processes or additional constraints beyond those used to characterize the feasible set may be relevant.

Our work provides the opportunity to more fully explore the feasible set approach by greatly decreasing computational time and by defining the feasible set of another macroecological pattern, i.e., the SSAD. Our computational and theoretical advances enabled us to examine combinations of constraint values that were previously out of reach. The algorithms we developed apply to frequency distributions such as the SAD and SSAD. However, many macroecological patterns are also cumulative, describing the rates at which species are encountered with increasing area (species-area relationship), time (species-time relationship) or both area and time (species-time-area relationship). Characterizing and randomly sampling the feasible sets of these and other patterns may require extensive modification of the algorithms we developed, approaches more similar to that of Haegeman & Loreau (2008), or altogether new approaches.

References

Amend, A.S., Seifert, K.A., Samson, R., & Bruns, T.D. (2010). Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *P. Natl.Acad. Sci. USA*., 107, 13748-13753.

Bóna, M. (2006). *A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory*. 2nd Edition. World Scientific Publishing Co., Singapore.

Brown, J. H. (1995). *Macroecology*. Univ. Chicago Press, Chicago.

Brown, J.H., Mehlman, D.W. & Stevens, G.C. (1995). Spatial variation in abundance. *Ecology*, 76, 2028-2043.

Chu, H., Fierer, N., Lauber, C.L., Caporaso, J.G., Knight, R. & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ*. *Microbiol*., 12,2998–3006.

Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A. & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J*., 6, 1007–17.

Flores, G.E., Campbell, J., Kirshtein, J., Meneghin, J., Podar, M., Steinberg, J.I. *et al*. (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol.*, 13, 2158-2171.

Haegeman, B. & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, 117, 1700–1710.

Harte, J., Zillio, T., Conlisk, E. & Smith. A.B. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology*, 89, 2700–2711.

McGlinn, D. J., & Hurlbert, A. H.. (2012). Scale dependence in species turnover reflects variance in species occupancy. *Ecology*, 93, 294–302.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al*. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386.

National Audubon Society. (2002). The Christmas Bird Count historical results. Retrieved from http://www.audubon.org/bird/cbc.

Nijenhuis, A. & Wilf, H.S. (1978). *Combinatorial Algorithms for Computers and Calculators*. Academic Press, New York.

Phillips, O. & Miller, J.S. (2002). *Global Patterns of Plant Diversity: Alwyn H. Gentry's Forest Transect Data Set*. Missouri Botanical Garden Press, St. Louis, Missouri, USA.

Sauer, J.R., Hines, J.E., Fallon, J.E., Parkieck, D.J., Ziolkowski, D.J. Jr. & Link, W.A. (2011). *The North American Breeding Bird Survey 1966-2009*. Version 3.23.2011. USGS Patuxent Wildlife Research Center, Laurel, MD.

Stojmenovic, I. (2008). Generating all and random instances of a combinatorial object. In *Handbook of Applied Algorithms: Solving scientific, Engineering and Practical Problems*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 1-38.

Storch, D., Šizling, A. L., Reif, J., Polechová, J., Šizlingová E., & Gaston, K.J. (2008). The quest for a null model for macroecological patterns: geometry of species distributions at multiple spatial scales. *Ecol. Lett.*, 11,771–784.

Supp, S.R., Xiao, X., Ernest S.K.M. & White E.P. (2012). An experimental test of the response of macroecological patterns to altered species interactions. *Ecology*, 93, 2505–2511.

Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and abundance of mammalian communities. *Ecology*, 92, 2316-2316.

U.S. Department of Agriculture, F.S. (2010). Forest inventory and analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis.

White, E. P., Thibault, K. M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

Table 3—1. Comparison of results from partitioning methods. Results of Locey and White (2013) and those from the reanalysis of those data used in that study reveal the practical gains of the algorithms developed here. Note, sampling effort per dataset was not controlled or accounted for.

Dataset	Number of SADs	Chapter 1 **~10K hours**	Present Study **~1K hours**
North American Breeding Bird Survey	2769	1586, 57%	2769, 100%
Christmas Bird Count	1992	129, 6.5%	1231, 62%
Gentry's forest transects	222	182, 82%	221, 99.5%
Forest Inventory and Analysis	10356	7359, 71%	10101, 98%
Mammal Community Database	103	42, 41%	103, 100%
Aquatic metagenomes	252	48, 19%	120, 48%
Terrestrial metagenomes	128	92, 72%	113, 88%
Fungi metagenomes	128	124, 97%	128, 100%
Total	15950	9562, **60%**	14786, **92.7%**

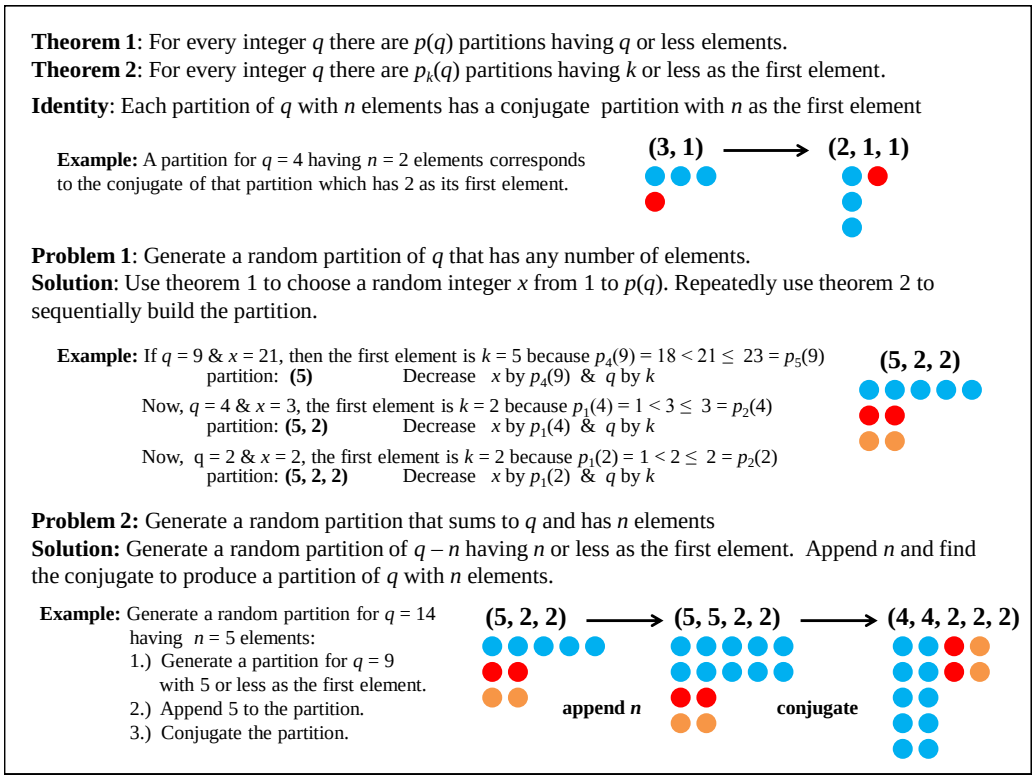


Figure 3—1. General approach for generating random integer partitions of a total having exactly a given number of parts. *Top:* Two theorems and one identity used to generate random integer partitions. *Center:* General method for generating a random partition of q having 1 to q elements using the two theorems. *Bottom:* General method for generating a random partition of q into exactly n elements, using the two theorems and the partitioning identity.

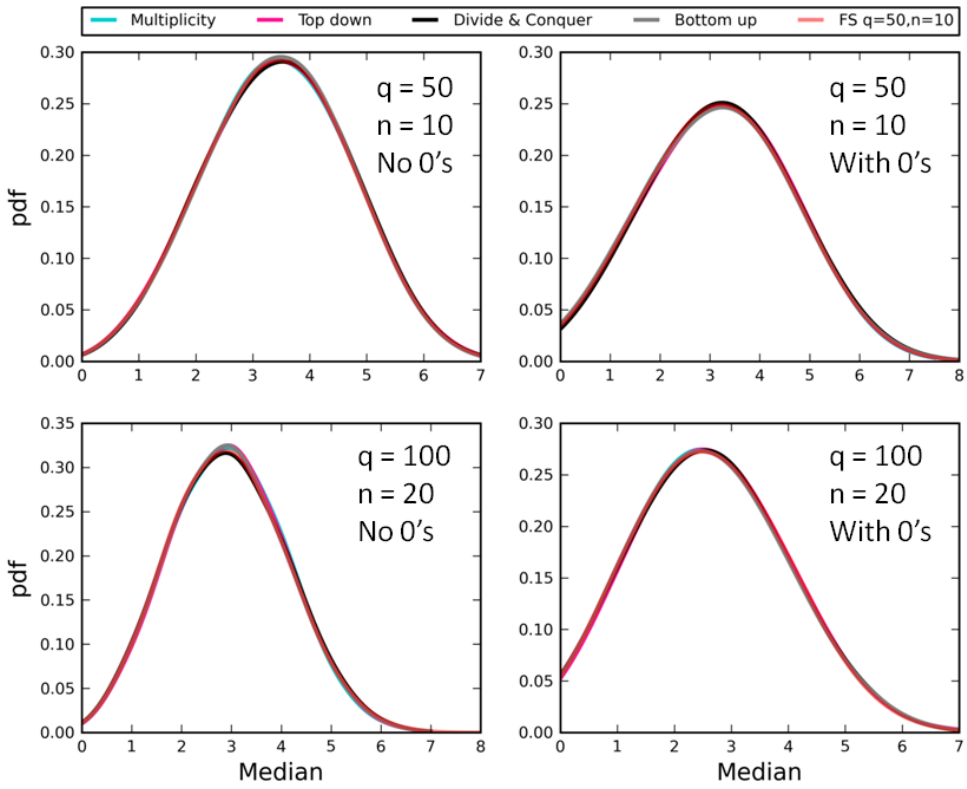


Figure 3—2. Kernel density curves revealing lack of bias in developed algorithms. A comparison of the full feasible set and kernel density curves for the median derived from 1000 random samples for different combinations of *q* and *n* using our four new algorithims for parts without zeros (left column) and for parts with zeros (right column). The similarity between the results derived using our algorithms and the full feasible set reveals that the algorithms produce unbiased random samples of the feasible set. We used Sage to generate the entire feasible set for $q = 50$ and $n = 10$ (16928 partitions) and used the random partitioning function in Sage to generate 1000 partitions for $q = 100$ and $n = 20$, which is too large to enumerate in full in reasonable time (10474462 partitions).

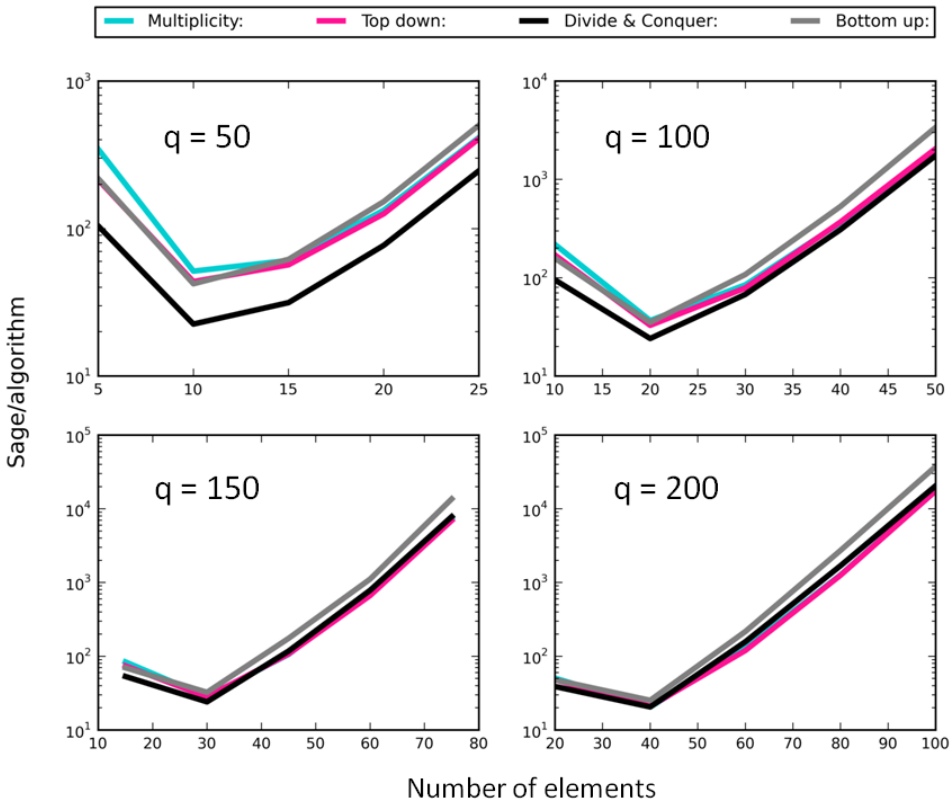


Figure 3—3. Comparison of speed between a pre-existing method and the methods derived in this chapter. Plots of the ratio of the computational time for Sage to generate 300 random integer partitions (no zeros) to the time taken for the new algorithms (‘Multiplicty’, ‘Top down’, ‘Divide and Conquer’, ‘Bottom up’) to do the same.

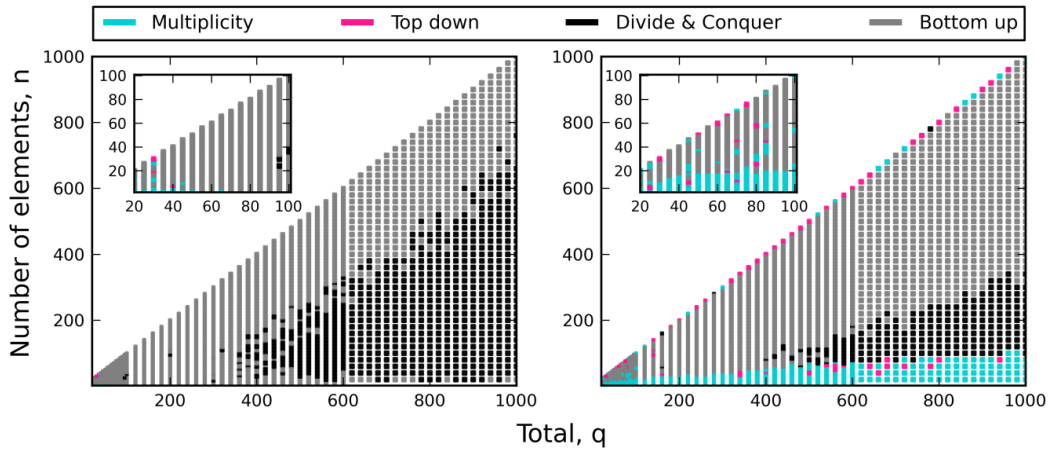


Figure 3—4. Color map revealing the fastest algorithm for specific combinations of $q \leq 1000$ and $n \leq q$. Comparisons were based on the time taken to generate 300 random partitions for each combination of q and n, both for cases where parts were allowed to have zero values (left) and when parts had positive values only (right). Insets reveal the small corner of the main graph where $q \leq 100$.

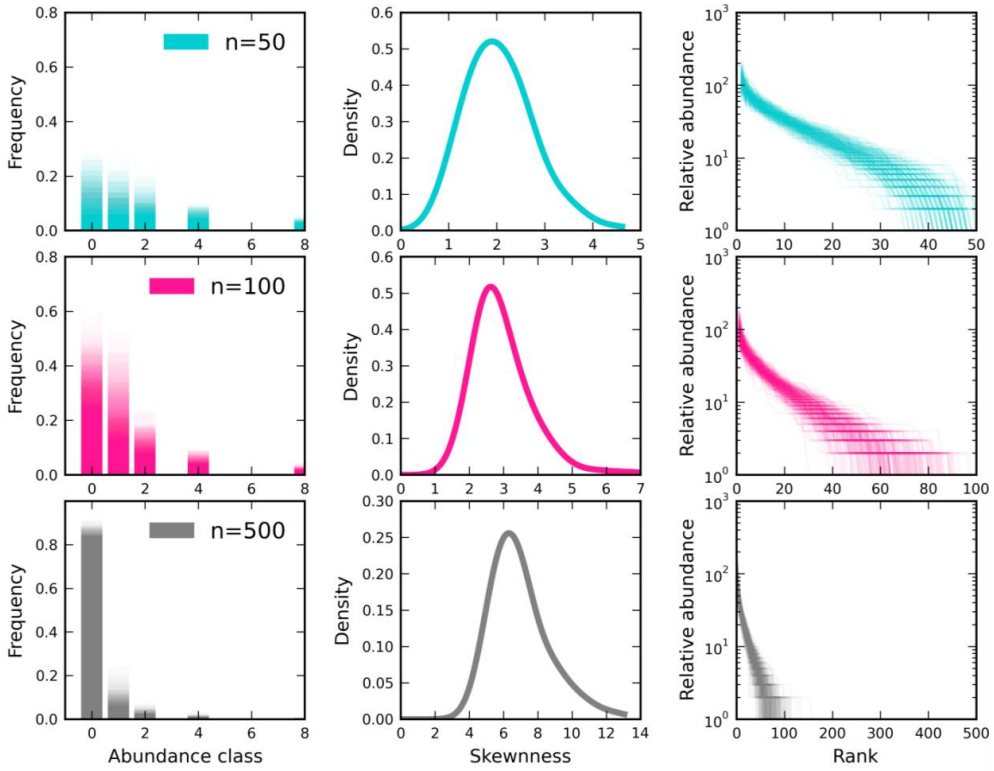


Figure 3—5. Plots revealing characteristics of the intraspecific spatial abundance distribution feasible set. The SSAD feasible sets for q = 1000 and n = {50, 100, 500} for 500 random samples displayed as frequency distributions (left), kernel density curve of skewness (center), and ranked distributions of abundance (i.e. areas or subplots ranked from most-to-least occupied) (right). Each row applies to one value of n. As q becomes distributed across an increasing number of elements, e.g. areas or quadrats, the feasible set in general becomes characterized by increasingly hollow-curves (i.e. few large elements and many small or zero-value elements), suggesting higher aggregation in SSADs as q is distributed across an increasingly large number of areas, quadrats, or subplots. For n = 500, ranks with values of 0 are plotted but difficult to discern.

CHAPTER 4

SIMPLE STRUCTURAL DIFFERENCES BETWEEN CODING AND NONCODING DNA*

Abstract

The study of large-scale genome structure has revealed patterns suggesting the influence of evolutionary constraints on genome evolution. However, the results of these studies can be difficult to interpret due to the conceptual complexity of the analyses. This makes it difficult to understand how observed statistical patterns relate to the physical distribution of genomic elements. We use a simpler and more intuitive approach to evaluate patterns of genome structure.

We used randomization tests based on Morisita's Index of aggregation to examine average differences in the distribution of purines and pyrimidines among coding and noncoding regions of 261 chromosomes from 223 microbial genomes representing 21 phylum level groups. Purines and pyrimidines were aggregated in the noncoding DNA of 86% of genomes, but were only aggregated in the coding regions of 52% of genomes. Coding and noncoding DNA differed in aggregation in 94% of genomes. Noncoding regions were more aggregated than coding regions in 91% of these genomes. Genome length appears to limit aggregation, but chromosome length does not. Chromosomes from the same species are similarly aggregated despite substantial differences in length. Aggregation differed among taxonomic groups, revealing support for a previously reported pattern relating genome structure to environmental conditions.

Our approach revealed several patterns of genome structure among different types of DNA, different chromosomes of the same genome, and among different taxonomic groups. Similarity in aggregation among chromosomes of varying length from the same genome suggests that individual chromosome structure has not evolved independently of the general constraints on genome structure as a whole. These patterns were detected using simple and readily interpretable methods commonly used in other areas of biology.

*Coauthored by: Locey, K.J. & White, E.P. (2011). Simple Structural Differences between Coding and Noncoding DNA. *PLoS ONE*, 6, 2, e14651.

Introduction

Evidence that selection affects the organization of information within genomes has resulted in efforts to characterize large-scale patterns of genome structure. Recently, advanced statistical and graphical methods such as chaos game theory, wavelet analyses, information theory, thermodynamics, and fractal geometry have been used to examine large-scale genome structure (Almeida & Vinga 2002, Garte 2004, Wang et al. 2004, Zhou et al. 2005, Allen et al. 2006, Oliver et al. 2008, Nandy 2009, Parker et al. 2009). The results of these studies have increased our knowledge of how genomes are organized by moving beyond simple characterizations such as genome length and GC content, to study how the distribution and organization of information within genomes may be evolutionarily constrained (Allen et al. 2006). While statistically informative, the structures quantified by these studies can be difficult to understand, making it difficult to interpret how the observed statistical patterns relate to the physical distribution of genomic elements.

Considering the difficulty of linking complex statistical patterns to the physical structure and biological processes affecting genomic evolution, we ask whether patterns in large-scale genomic structure can be quantified using a simpler approach with an intuitive structural interpretation. This simplification has the potential to allow for less statistically abstracted interpretations of genomic structural patterns. Here, we attempt such an approach using a straightforward definition of one of the most intuitive structural properties of sequential data, aggregation. We use this measure to detect a general difference among the two major kinds of DNA and the two forms of nitrogenous bases

commonly used in other studies (Almirantis & Provata 1997, Rogozin et al. 2002, Zhou et al. 2005, Allen et al. 2006, Oliver 2008, Nandy 2009, Parker et al. 2009).

Specifically, genomes are comprised of regions of DNA that code or do not code for proteins and are composed of two different structural forms of nitrogenous bases, purines (Pu) represented by adenine and guanine, and pyrimidines (Py) represented by thymine and cytosine. Assuming that coding and noncoding DNA are structured by different selective forces (Rogozin et al. 2002), common units of coding and noncoding regions (i.e. Pu and Py) may exhibit different distributions resulting from different structuring forces. Our aim was to use Morisita's Index of aggregation (I_M) (Morisita 1959, Morisita 1962, Morisita 1971, Hurlbert 1990) to examine whether: 1) Pu and Py exhibit non-random structure within sequences; 2) aggregation differs between coding and noncoding DNA; and 3) patterns of aggregation differ among chromosomes of the same species and among taxonomic groupings. If meaningful patterns can be detected this suggests that aggregation may provide an intuitive measure of structural genomic patterns that can be meaningfully influenced by biological processes.

Methods

Obtaining genomic data

We created Perl scripts to examine 261 chromosomes of 223 genomes from 21 phylum level microbial groups, downloaded from the National Center for Biotechnology Information microbial genome website, www.ncbi.nlm.nih.gov/genomes/lproks.cgi. We downloaded FASTA sequence and GenBank feature files. We picked genomes and

chromosomes that represented a broad range of lengths and protein coding contents. Pearl scripts and microbial genome information and per chromosome results can be accessed through here (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0014651).

Genome handling and aggregation estimation

We obtained estimates of aggregation for coding and noncoding DNA by using a sliding window approach to estimate the average aggregation of Pu and Py among consecutive non-overlapping 100-base sections of chromosomes. Rather than examine each individual coding or noncoding region separately, we examined coding and noncoding DNA as concatenated sequences of individual regions. These approaches alleviated two problems. First, analyzing individual coding and noncoding regions leaves a considerable amount of genome unanalyzed because individual coding and noncoding regions are rarely perfect multiples of a particular window size. Second, information regarding GC content is lost when sequences are binarily recoded according to Pu (A,G) and Py (C,T), hence removing potential statistical effects of GC content on aggregation.

We used Morisita's Index (I_M) (Morisita 1959, Morisita 1962, Morisita 1971, Hurlbert 1990) as our aggregation metric. I_M is commonly used in ecological and evolutionary studies (Williamson 1975, Ricklefs & Lau 1980, Sakai & Oden 1983, Dewey & Heywood 1988) to study the spatial distribution of age classes, genotypes, and species, and has been shown to be a more precise and less biased descriptor of spatial aggregation than other methods (e.g. variance:mean ratio) (Hurlbert 1990). I_M uses the number of occurrences among subsections of sampling areas (i.e., windows) to estimate measurements of aggregation based on a sampling probability. Specifically, I_M measures

how many times more likely it is that two randomly selected individuals will be from the same subsection of study area than if the individuals in the population were distributed at random. For example, I_M = 1.5 indicates that the probability of sampling two individuals from the same quadrat is 50% greater than if the population was randomly distributed (i.e., Poisson distributed). An I_M of 0.5 indicates this probability is 50% less likely than random. I_M is not typically used in cases of severely limited occupancy (e.g. linear segments of genomes of *n* size holding, at most, *n* Pu or Py). As a result, the value representing randomness was offset from I_M = 1.0 to I_M = 0.91936 (SE = .000057), as determined from 20,000 randomizations. Therefore we compared observed values to randomizations of the same sequence (see below) to determine if the genome was more of less aggregated than random and to determine whether or not this difference was statistical meaningful.

Morisita's Index is calculated as:

$$I_M = \left(\frac{X}{X-1}\right)\left(\frac{1}{\mu}\right)\left(\frac{\sigma^2}{\mu} + \mu - 1\right)$$

where X is the total number of individuals in the sampling universe, μ is the mean number of individuals per quadrat (i.e. subsection of the sampling universe), and σ^2 is the variance of individuals among quadrats. The formulation here is identical to that in Hurlbert (1990). In the present study, X is the total number of Pu (or Py) in a 100 base section of a genome, referred to here as a window, μ is the average number Pu or Py within each 10 base subsection of the window, and σ^2 is the variance of Pu or Py among

the 10 subsections. It can be seen from the above equation that Morisita's Index is independent of genome length, genome segment length, and number of genome segments and is thus independent of the density of individuals in the window (Hurlbert 1990). Using I_M thus controlled for differences in the density of Pu and Py among genomes. We also confirmed that I_M was insensitive to window and subsection size by reanalyzing a random subset of 29 genomes using several combinations of window size (100, 400) and subsection size (10, 20, 40). These combinations yielded qualitatively similar results (see table in supplementary materials).

Randomizations

We created 100 randomized versions of each genome for comparison with actual genomes by randomly redistributing Pu and Py within individual coding and noncoding regions. These randomized genomes were analyzed as described above for comparison to actual genomes. By avoiding changes in the number of Pu and Py among individual regions, observed differences reflect the effect of nitrogenous base order; another control for the effects of Pu and Py density. P-values were determined to be less than 0.01 when average measurements of I_M from real genomes were greater than those from all 100 randomizations or less than those from all 100 randomizations.

Statistical analysis

Microbial genomes typically contain a much larger fraction of coding than noncoding DNA. Here, the percentage of coding DNA ranged from 73.54 to 95.54%. Under this circumstance, I_M is calculated more times for coding DNA (typically tens of

thousands) than noncoding DNA (typically thousands). To account for this difference in sample size, we chose non-parametric rank-sum tests to determine whether Pu and Py generally differ in aggregation between coding and noncoding regions of individual genomes. Additionally, we conducted Spearman's rank correlation to determine whether aggregation was related to percent coding DNA, genome length, chromosome length, and GC-content. We chose a nonparametric correlation technique because all datasets were non-normally distributed as determined from the Lilliefors test for normality. We used the student version of MATLAB v7.7.0 to generate kernel density curves, box plots, and to conduct all statistical analyses.

Results

Purines (Pu) and pyrimidines (Py) were distributed similarly within genomes and chromosomes, as illustrated by nearly identical distributions within coding and noncoding DNA (Fig. 1) and the similar results of statistical analyses (Table 4—1 and 4—2). In coding DNA Pu and Py were less aggregated (i.e. more evenly distributed) than random in approximately 44% of genomes, and more aggregated than random in almost 52% of genomes ($p < 0.01$; Table 4—1). Noncoding DNA was rarely more evenly distributed than random (~10% of genomes) with 86% of genomes exhibiting significant aggregation ($p < 0.01$; Table 4—1). The difference in aggregation between coding and noncoding DNA was significant in 94% of chromosomes ($n = 245$). Of these 245 chromosomes, noncoding DNA was more aggregated than coding DNA in 91% of cases ($n = 224$). Hence, coding DNA was more aggregated than noncoding DNA in only 21 chromosomes (8.0%), from 18 genomes.

Of the 18 genomes (21 chromosomes) where coding DNA was more aggregated than noncoding DNA, seven genomes belong to the Spirochaetes group. The other 11 genomes are widely distributed across groups: Alphaproteobacteria (3), Aquificae (1), Bacterioides/Chloribi (1), Betaproteobacteria (1), Crenarcheota (2), Euryarchaeota (1), Gammaproteobacteria (1), and Nanoarcheota (1). Only two of the 13 Spirochaete members represented in the dataset showed greater average aggregation in noncoding DNA than coding DNA. Compare this to Actinobacteria (N = 17), Thermotogae (N = 8), Firmicutes (N = 15), and Epsilonproteobacteria (N = 9) where all members showed greater average aggregation in noncoding DNA, or to Gammaproteobacteria (N = 32), Euryarchaeota (N = 11), or Betaproteobacteria (N = 26) where all but one member showed greater average aggregation in noncoding DNA. All other groups had three or fewer members lacking greater average aggregation within noncoding DNA than coding DNA. Hence, Spirochaetes appear to be the only phylum-level group where noncoding DNA is not typically more aggregated than coding DNA.

Aggregation varied significantly among phyla, with individual groups of taxa typically occupying narrow ranges of aggregation and having little-to-no overlap with most other groups (Fig 4—2). However, the distribution of taxonomic groups across the observed range of aggregation revealed no apparent phylogenetic clustering or pattern. For instance, proteobacteria are distributed throughout while archaeal groups are separated by bacterial groups. When the set of 200 genomes was examined as a group, with an average measure of aggregation for each genome represented by a single data point, coding and noncoding regions formed different distributions of aggregation with

noncoding regions shifted towards higher values of aggregation (Fig. 4—1). Despite a smooth unimodal distribution of aggregation values among noncoding DNA, coding DNA from the identical set of genomes exhibited an apparent bimodality. While the first mode could be the result of sample bias, the lack of a corresponding mode in the curve for noncoding DNA suggests two different subgroups of genomes with aggregated noncoding DNA; one where the distribution of nitrogenous bases in coding DNA is under-aggregated to essentially random ($I_M = 0.91936$), and one where the distribution is significantly aggregated.

Aggregation, as estimated with Morisita's Index, showed a significant correlation with GC content and a slight but also significant correlation with percent coding DNA (Fig. 4—3). Aggregation was also significantly correlated with genome length. The strength of the correlation and the shape of the distribution reveals that estimates of I_M decreased and converged on lower values with increasing genome length (Fig. 4—4), suggesting that larger genomes tend to be less aggregated. Among genomes with multiple chromosomes, I_M and chromosome length were not correlated (Fig. 4—4). However, when the lengths of these chromosomes were summed to obtain the length of the genome, the pattern of limited aggregation with increasing genome length was again obtained (Fig. 4—2). Additionally, aggregation was similar among chromosomes of the same species (average % difference = 0.28 ± 0.04 SE for Py to 0.27 ± 0.04 SE for Pu) despite large differences in chromosome length (average % difference = 91.3 ± 9.23 SE).

Discussion

Both structural forms of nitrogenous bases clearly exhibit non-random distributions within genomic sequences and are nearly always distributed similarly. Steps taken to remove statistical effects of density, sampling scale, and GC bias, and to examine the statistical relationships of aggregation to GC content, percent coding DNA, and genome length reveal that the reported differences between coding and noncoding DNA are likely due to meaningful patterns of Pu and Py clustering within sequences and not due to the statistical effects of these other variables. Despite removing statistical effects of GC-content by recoding genomes in Purines and Pyrimidines, and using a measure of aggregation that is independent of the ratio of coding to noncoding DNA, GC-content, genome length (but not chromosome length), and percent coding DNA were significantly correlated with aggregation. Though these results suggest that relationships among these structural genomic features are real, further studies will be necessary to understand these patterns.

Genome length appears to set a maximum limit on the degree of aggregation possible (Fig. 4—4). This pattern holds for genomes with single and multiple chromosomes. However, the lengths of chromosomes from multi-chromosomal genomes do not appear to show the same relationship. Instead, chromosomes of the same species are similarly aggregated despite large differences in length. When the lengths of these chromosomes are summed to obtain overall genome length, their summed lengths follow the decreasing pattern shown for single chromosome genomes (Fig. 4—4). At the chromosome scale, aggregation appears to be a property of the species, largely invariant

with chromosome length. However, overall aggregation seems to be limited by genome length, perhaps regardless of the number of chromosomes comprising a genome. Both similarity in aggregation among chromosomes of varied length from the same genome, and the tendency for aggregation among chromosomes to be influenced by overall genome length, suggests that chromosome structure has not evolved independently of general constraints on overall genome structure.

Noncoding DNA was almost always more aggregated than coding DNA. In other words, nitrogenous bases of similar structure are more likely to be found in close proximity within noncoding DNA than within coding DNA. This conclusion is based on the genome-wide averaging of tens of thousands of estimates of I_M across a diverse collection of 223 microbial genomes, and hence, represents a general low-resolution pattern of genome structure. It may be unlikely that such a pattern is the result of one or even a few specific genetic or evolutionary processes. What it does suggest is that the functions that coding and noncoding DNA perform, and the pressures that affect their evolution, are different enough to manifest a general difference in the gross distribution of their common elements.

For Spirochaetes, the pattern is typically reversed. Spirochaetes are a small and cohesive group of gram-negative chemoheterotrophs. They are unusual in their linear chromosomes, cytoskeleton, long helical cells, and coevolution with a host-specific phage. As such, it is possible that these traits that distinguish Spirochaetes from other microbes explain their exception to the general pattern. However, a superficial investigation of the microbial traits is unlikely to explain this reversed pattern, because a

variety of cell shapes (e.g. coccus, rod, spiral), chromosome shapes (e.g. linear, circular), temperature ranges (e.g. mesophilic, thermophilic), habitats (e.g. soils, sulfur springs, hosts), chromosome lengths (490885-5566749), and percent coding DNA (0.7475-0.9483), are represented within the set of 18 genomes where coding DNA was on average more aggregated that noncoding DNA.

The observed bimodality in the distribution of aggregation values for coding DNA suggests the presence of two general groups of genomes differing characteristically in the patterns of aggregation within coding DNA. Whether these two groups differ in a biologically meaningful way that influenced the distribution of structurally different nitrogenous bases has not yet been determined. Further investigation is necessary to determine whether this bimodality results from the specific genomes chosen for analysis or whether it is an indicator of an important biological process that has shaped genome evolution among microbes.

The distribution of phyla across the range of aggregation in this study strongly corroborates the pattern described by Bohlin et al. (2009) who examined the genomic fraction of purine and purine/pyrimidine stretches (i.e. an indirect measure of aggregation) in relation to environmental variables across a similar but smaller set of prokaryote phyla (Bohlin et al. 2009). Though there are no methodological similarities, and noncoding DNA is analyzed separately from coding DNA in this study, both studies reveal that phyla occupy similarly ordered and narrow ranges of aggregation (Table 3). When comparing the ranks of phyla common to both studies there were four exact matches and four instances where phyla differed by only one rank. The reproduction of

this pattern in spite of minimal methodological similarity suggests that the pattern is robust and relatable to functional traits that interface with the exogenous environment (Bohlin et al. 2009).

Despite the potential for exceedingly complex distributions of bases within coding and noncoding regions, the study of large-scale genomic structure clearly does not preclude the use of simple approaches to arrive at general patterns based on intuitive properties. It is clear that those forces that have structured protein coding and noncoding regions, as well as individual chromosome and overall genome structure, have left evidence of their effects at the level of common elements, the two types of structural nitrogenous bases. We suggest that processes and constraints with predominant effects on genome structure should influence the patterns of aggregation observed in this study. While statistical approaches to large scale genome structure have the potential to reveal novel and meaningful patterns as well as structural relationships, we suspect that the general patterns reported here are unlikely to be explained by statistical approaches alone, that is, without establishing the genetic or evolutionary mechanisms. Lack of clarity in the interpretation of statistical methods, metrics, and results that document novel and poorly understood structural patterns can only be a detriment to this endeavor.

References

Allen, T.E. , Price, N.D. , Joyce, A.R. & Palsson, B.Ø. (2006). Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput. Biol.*, 2, e2.

Almeida, J.S. & Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics,* 2002, 3.

Almirantis, Y. & Provata, A. (1997). The “clustered structure” of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. *B. Math. Biol.*, 59, 975-992.

Bohlin, J., Hardy, S.P. & Ussery, D.W. (2009). Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes. *BMC Genomics*, 2009, 10, 346.

Cristea, P.D. (2002). Conversion of nucleotides sequences into genomic signals. *J. Cell Mol. Med.*, 6, 279-303.

Dewey, S.E. & Heywood, J.S. (1988). Spatial genetic structure in a population of *Psychotria nervosa*. I. Distribution of genotypes. *Evolution*, 42, 834-838.

Garte, S. (2004). Fractal properties of the human genome. *J. Theor. Biol.*, 230, 251-260.

Hurlbert, S.H. (1990). Spatial distribution of the montane unicorn. *Oikos*, 58, 257-271.

Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., *et al*. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17, 149-154.

Marenduzzo, D., Micheletti, C. & Cook, P.R. (2006). Entropy-drive genome organization. *Biophys. J*., 90, 3712-3721.

Mitra, A., Liu, G. & Song, J. (2009). A genome-wide analysis of array-based comparative genomic hybridization (CGH) data to detect intra-species variations and evolutionary relationships. *PLoS ONE*, 4, e7978.

Morisita, M. (1959). Measuring of the dispersion of individuals and analysis of the distributional patterns. *Mem. Faculty Sci, Kyushu Univ. Ser. F*. (*Biol*), 2, 215-235.

Morisita, M. (1962). I-index, a measure of dispersion of individuals. *Res. Popul. Ecol*., 4, 1-7.

Morisita, M. (1971). Composition of the I-index. *Res. Popul. Ecol.*, 13, 1-27.

Nandy, A. (2009). Empirical relationships between intra-purine and intra-pyrimidine difference in conserved gene sequences. *PLoS ONE*, 4, e6829.

Oliver, J.L., Bernaola-Galván, P., Hackenberg, M. & Carpena, P. (2008). Phylogenetic distribution of large-scale genome patchiness. *BMC Evol Biol*., 8, 107.

Parker, S.C.J., Hansen, L., Abaan, H.O., Tullius, T.D. & Margulies, E.H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, 324, 389-392.

Ricklefs, R.E. & Lau, M. (1980). Bias and dispersion of overlap indices: results of some monte carlo simulations. *Ecology*, 61, 1019-1024.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., *et al* (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic*. *Acid*. *Res*., 30, 4264-4271.

Sakai, A.K. & Oden, N.L. (1983). Spatial pattern of sex expression in Silver Maple (*Acer saccharinum* L.): Morisita's Index and spatial autocorrelation. *Am*. *Nat*., 122, 489-508.

Wang, Y., Hill, K., Singh, S. & Kari, L. (2004). The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*, 346, 173-185.

Williamson, G.B. (1975). Pattern and Seral Composition in an Old-growth Beech-Maple forest. *Ecology*, 56, 727-731.

Zhou L., Yu, Z., Deng, J., Anh, V. & Long, S. (2005). A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation. *J*. *Theor*. *Biol*., 232, 559-567.

Table 4—1. Aggregation among microbial genomes.

Genomes, N = 223	Coding		Noncoding	
	Pu	Py	Pu	Py
Aggregated	52.0% (n = 116)	52.0% (n = 116)	86.1% (n = 192)	86.1% (n = 192)
Random	5.4% (n = 12)	4.0% (n = 9)	3.6% (n = 8)	4.0% (n = 9)
Overdispersed	42.6% (n = 95)	44.0% (n = 98)	10.3% (n = 23)	9.9% (n= 22)

Table 4—2. Aggregation among microbial chromosomes.

Chromosomes N = 261	Coding		Noncoding	
	Pu	Py	Pu	Py
Aggregated	46.7% (n = 122)	47.9% (n = 125)	80.4% (n = 210)	80.5% (n = 210)
Random	5.4% (n = 14)	5.4% (n = 14)	5.4% (n = 14)	5.7% (n = 15)
Overdispersed	47.9% (n = 125)	46.7% (n = 122)	14.2% (n = 37)	13.8% (n= 36)

Table 4—3. Phyla ranked according to aggregation of purines, averaged for coding and noncoding DNA, as reported here and as reported by Bohlin et al. (2009). Ranks in parentheses (n = 4) are exact matches, ranks with asterisks (n = 4) are one rank different.

	Present Study	Bohlin et al. (2009)
Rank	Purine Aggregation	Purine Stretches
1	Chlamydia	Thermotoga
2*	Thermotoga	Spirochaetes
3	Firmicutes	Chlamydia
4	Spirochaetes	Euryarcheota
5	Deltaproteo	Crenarchaeota
6*	Crenarchaeota	Firmicutes
(7)	Epsilonbacteria	Epsilonbacteria
8*	Cyanobacteria	Deltaproteo
9	Alphaproteo	Cyanobacteria
(10)	Gammaproteo	Gammaproteo
11	Euryarcheota	Chloroflexi
12*	Chloroflexi	Alphaproteo
(13)	Actinobacteria	Actinobacteria
(14)	Betaproteo	Betaproteo

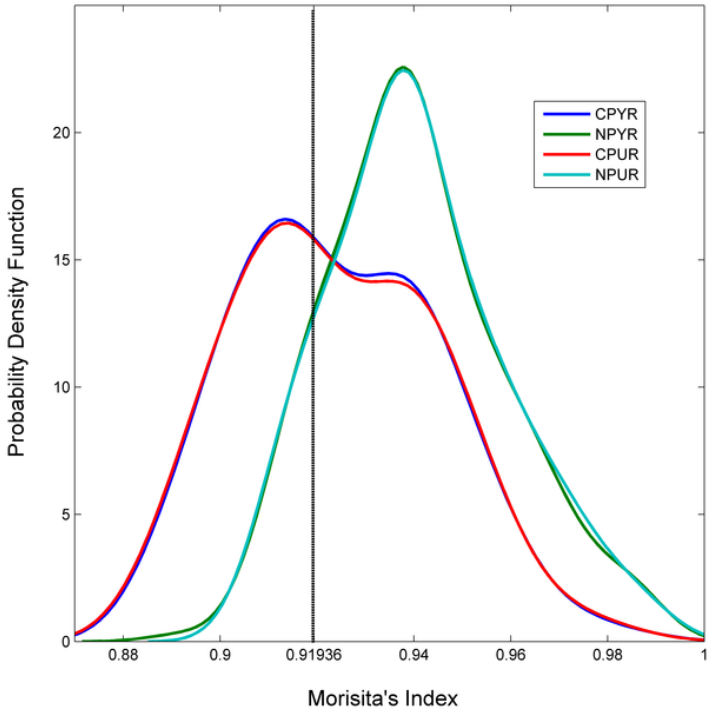


Figure 4—1. Kernel density curves reveal different distributions for coding and noncoding DNA. Kernel density curves for purines and pyrimidines within coding (C) and noncoding DNA (N). Distributions for purines and pyrimidines nearly completely overlap. Curves for noncoding DNA are shifted towards higher values of aggregation while curves for coding DNA are centered closer to the derived value for randomness, 0.91936. Apparent bimodality within coding regions may have resulted from the sample-size of different taxonomic groupings (e.g. 32 Gammaproteobacteria within a narrow range), but note the lack of bimodality among corresponding noncoding regions of the same set of genomes.

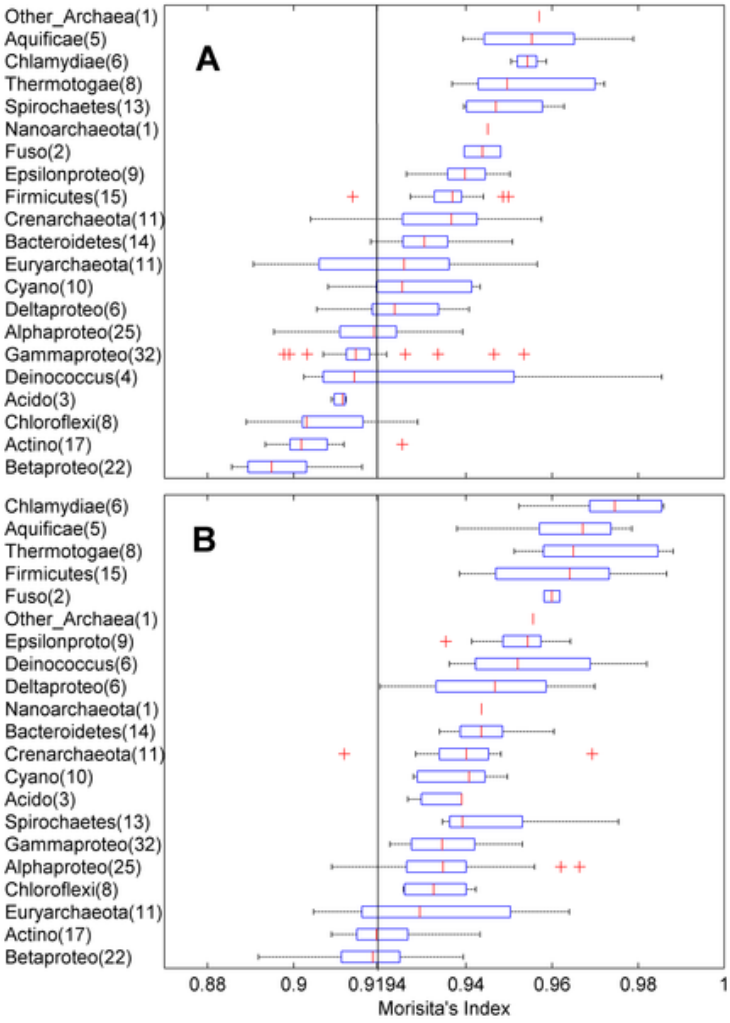


Figure 4—2. Box plots showing ranges of aggregation values (I_M) for pyrimidines within coding and noncoding DNA of 21 microbial groups. The distribution of box plots for coding DNA (A) is shifted more towards lower values of aggregation and closer to randomness than those for noncoding DNA (B) which are shifted towards values of higher aggregation.

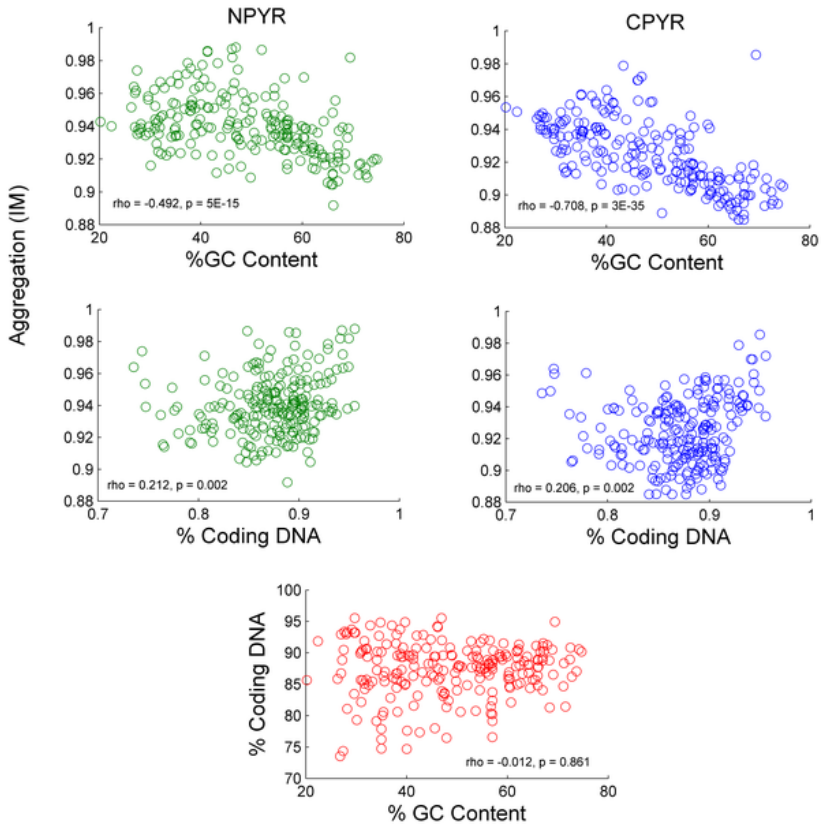


Figure 4—3. Plots of aggregation (I_M) vs. % GC content and % coding DNA, with a plot of % coding DNA vs. % GC content. Aggregation of pyrimidines within coding DNA (blue) and noncoding DNA (green) shows a greater linear relationship to %GC content than to % Coding DNA. % Coding DNA and % GC (red) content are not correlated.

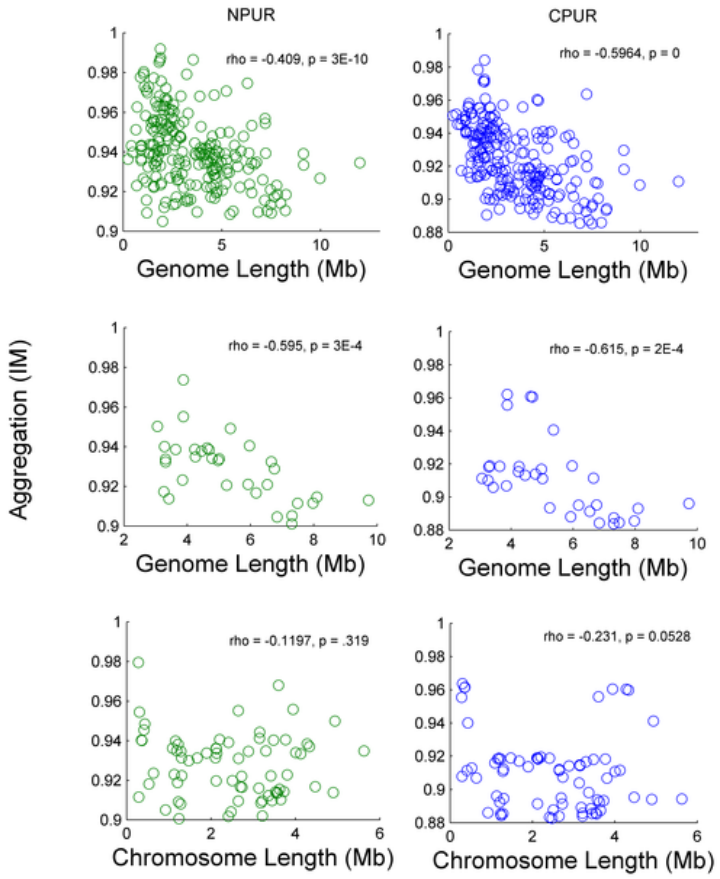


Figure 4—4. Plots of aggregation (I_M) vs. genome length and chromosome length for Purines (Pu). (Top) Aggregation of purines in coding (blue plots) and noncoding (green plots) DNA for 223 genomes. (Middle) Aggregation of purines in coding and noncoding DNA for the 33 genomes with multiple chromosomes. (Bottom) Aggregation of purines in coding and noncoding DNA for 71 individual chromosomes from the 33 genomes with multiple chromosomes. These plots reveal that dissecting a genome into its constituent chromosomes destroys the generally decreasing pattern of aggregation with increasing genome length.

CHAPTER 5

CONCLUSION

General variables related to the size of a system and the numbers of elements within it (e.g. size of a community, numbers of species, length of a genome) constrain structural properties of abundance and aggregation across systems representing disparate scales of biology, i.e. individual genomes and ecological communities. Though understanding the mechanisms that drive the structure of biological systems is a primary goal of biology, understanding ways in which the structures of biological systems are inherently constrained by general variables is, perhaps, of equal importance. Indeed, from the shapes of molecules to patterns of abundance in ecological communities, the structure of all biological systems is likely to be constrained by such variables as the size of the system and the numbers and types of components, parts, and elements within it.

In part, I revealed how genome length constrains the aggregation of nucleotides and how the aggregation of nucleotides differs between regions of DNA that either code or do not code for proteins. However, reasons for why genome length would constrain nucleotide aggregation remain unknown. Likewise, in this research (i.e. chapter four), I related nucleotide aggregation to the genetic information carried on the DNA molecule. However, this was with respect to protein coding information. It is well known that some noncoding DNA is transcribed into noncoding RNA elements (transfer RNA, ribosomal RNA). Consequently, not all noncoding DNA is without transcribable information.

Aggregation of nucleotides may also relate to the physical properties of DNA and the different physical or structural (i.e. non-informational) functions that noncoding DNA

performs. One way to begin teasing apart whether differences in nucleotide aggregation between coding and noncoding DNA actually relate to informational or structural properties would be to examine regions of noncoding DNA that either provide informational functions (e.g. regulation, methylation) or are transcribed into noncoding RNA elements, separately from regions of noncoding DNA that serve a non-informational but structural function, such as spatial separation of genes that lessens the influence of mutation in a specific region of a chromosome or that help bind the DNA molecule and prevent damages from physical, chemical, or thermal stress.

Once again, whether differences in nucleotide aggregation between coding and noncoding DNA are primarily due to informational or purely structural properties, the question might still remain as to why genome length would effectively place an upper bound on nucleotide aggregation where larger genomes, but not larger chromosomes, lead to lower maximum aggregation. These are the next most obvious steps in understanding why aggregation, a general statistical property, differs so clearly between coding and noncoding DNA. Likewise, there are several promising and important future directions to better understand the influence of general constraints on properties of aggregation and structure in ecological systems, specifically through the use of the feasible set approach.

In my initial feasible set work, i.e. chapters 2 and 3, I laid the groundwork for examining the distribution of abundance among species based solely on the shapes of the distribution that were possible given the values of total abundance (i.e. the number of individuals in a community) and species richness (i.e. the number of species in a

community). This approach was only concerned with the observable variation in the shape of the distribution, and it assumed nothing of how any particular shape may arise. Hence, this initial feasible set work took an entirely ‘macrostate’ based approach, i.e., an approach that considers only the shape of a distribution and not how the distribution may arise nor which values of abundance different species may have. For example, in macrostate form, the distribution of abundance among species only reveals that there are, say, two species with 10 and 20 individuals, but does not consider which species has 10 and which has 20. This later approach of considering the form of a distribution with respect to which entities (e.g. species, subplots) have which values (e.g. 10 or 20 individuals), is commonly referred to as a ‘microstate’ based approach.

It is possible to explicitly consider the feasible set of microstates, i.e. *ordered* ways in which the abundance of species (or subplots, etc.) can sum to a given total abundance. Just as there is a combinatorial object (i.e. integer partitions) for macrostates, there is a combinatorial object for microstates (i.e. integer compositions). Indeed, integer compositions are simply permutations of integer partitions. Because of this, the microstate feasible set based on integer compositions (*ordered* configurations of *unlabeled* individuals) can be many times larger than the macrostate feasible set. This microstate feasible set has yet to be explored or compared to empirical distributions of abundance, though it may well be that for some biological systems, the microstate feasible set explains more variation in abundance among species, subplots, samples, alleles, etc. than the macrostate feasible set. Likewise, there are other ways of considering or defining microstates (e.g. *ordered* configurations of *labeled* individuals among

samples or subplots) as well as different feasible sets based on the different possible microstates. But, perhaps most importantly when considering macrostate and microstate based configurations of ecological or biological patterns of abundance and aggregation, as well as their respective feasible sets, little work has been conducted to reveal when and which forms of microstates or macrostates are most appropriate for a given ecological or biological system or pattern.

In general, the feasible set approach should apply whether considering macrostates and observable variation or the potentially different forms of microstates and the ways in which different macrostates can arise. Most importantly, we should recognize that there may be a tenable feasible set, i.e., a set of alternative outcomes that can be examined, regardless of the 'state' with which we are working. When this is true, it implies the potential to examine and understand patterns of aggregation and abundance within biological systems solely with respect to a set of possible alternative outcomes, and hence, to understand how the structures of biological systems are constrained by and explainable from the values of numerical constraints. Though numerical constraints themselves may not be the focus a study, understanding patterns of aggregation and abundance in biological systems first in terms of known information (i.e. constraint values) reveals whether there is sufficient variation in the possible shapes of a pattern or set of outcomes to interpret the form of a pattern or outcome in terms of more complex phenomena such as biological, ecological, and statistical processes and mechanisms.

There are a great number of biological systems between the scales of individual genomes and ecological communities where patterns of aggregation and abundance are

studied. Examples are the distribution of frequencies among alleles, aggregation in occupancy of space or occurrence in time of individuals and species, branching patterns in phylogenetic trees, the partitioning of energy among an organism's metabolic processes, coalescent pathways of individual relatedness and metabolic pathways of cellular respiration. These systems are all characterized by a number of components (i.e. individual or group level entities) that interact in particular ways. Likewise, the structure of these systems, i.e. varying configurations of elements resulting in patterns of aggregation and abundance, may well be constrained by a feasible set determined by the values of such general variables as the number of individual organisms, species, alleles, taxa, subplots in the system.

Finally, research into the constraining influences of general variables on the structure of biological systems may represent one path towards synthesis between systems biology and macroecology. Both disciplines represent a large-scale holistic perspective on biological (system biology) and ecological (macroecology) systems; as opposed to a more proximate and experimental reductionist approach to identifying process and mechanism. Both systems biology and macroecology are driven by the study of emergent properties and the extensive use of theory, analytic and computational modeling, and highly quantitative approaches. However, while systems biology still makes use of experimental approaches, macroecology is largely pursued through the statistical analysis of patterns in abundance, distribution, and diversity. However, realization that the structure of both biological and, more specifically, ecological systems may be strictly constrained by general variables (e.g. size of the system, numbers and

types of parts, components, and elements) suggests room for greater synthesis and crossover between macroecology and systems biology.

APPENDICES

APPENDIX A:

REPRINT POLICIES FOR DISSERTATION CHAPTERS

Chapter 2—As per the publisher's (Wiley-Blackwell) author reprint policy: (http://authorservices.wiley.com/bauthor/faqs_copyright.asp#1.7), authors maintain the right to reprint manuscripts in theses/dissertations. No request from Wiley-Blackwell to reprint the work from Chapter 2 is needed.

Chapter 3—Is a preprint (not published) and requires no reprint permissions.

Chapter 4—Was published in PLoS ONE (http://www.plosone.org/). PLoS ONE maintains a Creative Commons Attribution License (CCAL), "Under the CCAL, authors retain ownership of the copyright for their article, but authors allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in PLOS journals, so long as the original authors and source are cited. No permission is required from the authors or the publishers." see http://www.plosone.org/static/license.

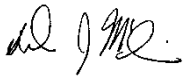
APPENDIX B:
CHAPTER 3: LETTER OF RELEASE FROM DR. DANIEL J. MCGLINN

13 December 2013

TO WHOM IT MAY CONCERN:

I hearby release the rights to reproduce the manuscript, "Efficient algorithms for sampling feasible sets of macroecological patterns" to my coauthor Kenneth J. Locey for use in his dissertation.

Daniel J. McGlinn



APPENDIX C:

CHAPTER 2: SUPPLEMENTAL FIGURES AND DETAILS ON DATASETS

Additional details on datasets—We used a subset of the data compiled by White et al. (2012); details provided in Appendix A of that paper. Consequently, much of the following is a reiteration of White et al.'s (2012) appendix. Some of the language is directly quoted from White et al. because it is the most precise way to convey the information. We used as much of the data in White et al. (2012) as was practical given the computationally intensive nature of our approach, which sometimes required several minutes to generate a single random macrostate for some combinations of total abundance (N) and species richness (S). Consequently, there were some combinations of N and S for which few, if any, macrostates could be generated. Details on each of those datasets, as well as on our use microbial metagenomic data are given below.

Birds—White et al. (2012) compiled data collected in 2009 from the North American Breeding Bird Survey (BBS; Sauer et al. 2011) and from the Christmas Bird Count (CBC; Christmas Bird Count (National Audubon Society 2002). BBS routes are 40 km long and consist of 50 three minute point counts, 800 m apart, sampled annually in June. CBC count circles are 24.1 km in diameter and are censused by multiple observers over the course of a day. Nocturnal species, water birds, and raptors were excluded from analyses, because they are poorly sampled by Breeding Bird Survey survey methods.

Trees—We used data compiled by White et al. (2012) from the USFS Forest Inventory Analysis program (U.S. Department of Agriculture 2010, Woudenberg et al. 2010,) (FIA), and the Alwyn H. Gentry Forest Transect Data Set (Phillips and Miller 2002) (GENTRY). We used one year of data (calendar year of sampling varies among plots) for FIA phase 2 plots that were sampled using the standardized methodology implemented in 1999 [see the FIA National Core Field Guide for more information (U.S. Department of Agriculture 2010)]. The standard plot consists of four 24.0-foot (7.32 m) radius subplots, on which trees 5.0 inches (12.7 cm) and greater in diameter are identified to species and measured. We further restricted our analysis of FIA data beyond the subsetting done by White et al. to include only natural forest stands (e.g. absence of human disturbance, plots without artificial regeneration, plots without silvicuture treatment). In addition we fixed a minor bug in the White et al. (2012) subsetting (see https://github.com/weecology/white-etal-2012-ecology/commit/90896ebf745341d7c8bbedb81d391b22c2673d87), which results in minor differences in some of the plots and years included in the analysis.

White et al. (2012) used data from the GENTRY dataset for 226-0.1 hectare sites throughout the world, with each site sampled once over the course of a 22 year period. The following is largely a reiteration of the data presented in the appendix of that study. At each site, all plants with stem diameters of 2.5 cm or greater were identified and measured along ten 2 × 50 m transects. Due to difficulties in the taxonomy and identification of tropical trees, some species in the Gentry dataset are identified only as morpho-species (unique within sites), and species' names vary among sites due to both

typographical errors and synonymy problems. Since we only analyzed data within a site, these issues do not affect our analyses[

Mammals—White et al. (2012) used species abundance data for the 42 sites included in the Mammal Community Database (Thibault et al. 2011) (MCDB). These mammal data were compiled from published sources and therefore were not collected using a standardized protocol across sites and therefore include various levels of sampling effort spread across varying amounts of time and space.

Microorganisms—Whereas our macroorganismal datasets (birds, trees, mammals) report abundances in numbers of individuals, species and abundance counts for microorganisms cannot be conducted *in situ* at the level of the individual cell. Instead, microbial ecologists rely on environmental sampling and molecular methods to sequence the DNA from an environmental sample to 1.) identify microbial taxa in the environment (by comparing cloned sequences to libraries of annotated sequences) and 2.) to calculate estimates of taxa abundance in the environment from estimates of abundance in the sample (using sophisticated alignment and clustering algorithms). Microbial taxa are often delineated based on percent rRNA or ITS (internal transcribed spacer) sequence similarity among sequences cloned from a sample. In the present paper, we used the 97% sequence similarity convention but also show that our results remain the same for 95% and 99% similarity cutoffs (see Appendix C). We obtained all of our metagenomic data from the MG-RAST metagenome server. MG-RAST reports taxa abundances as the estimate of the number of sequences in the sample that contain a given annotation. For

our study, an annotation pertains to an rRNA or ITS sequence belonging to a specific taxon.

Figure 1. Kernel density curves for evenness (E_{var}) across entire feasible sets (red) and samples of randomly drawn macrostates (black). Each subplot contains 10 kernel density curves based on 300, 500, and 700 random samples. These plots reveal that 300 random macrostates capture the structure of the feasible set nearly as well, or as well as, samples size of 500 and 700.

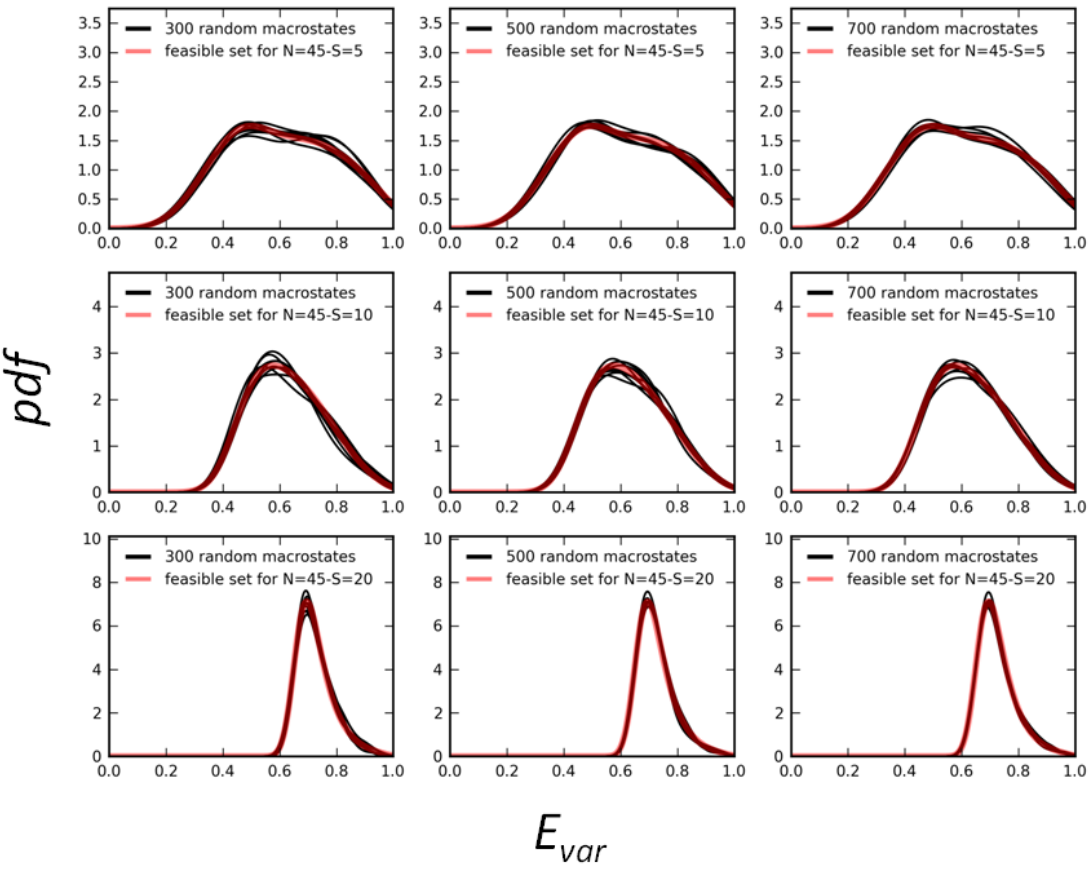


Figure 2. Heat maps of rank-abundance distributions generated from 500 randomly drawn macrostates from the feasible sets of N = 500 and S = {50, 100, 200, 400}. Heat maps, i.e. three-dimensional histograms of log(abundance) vs. rank values, effectively reveal the density of overlap among the 500 randomly drawn macrostates. The macrostate derived as the central tendency (grey line) runs through the densest regions (i.e. hottest regions) of the feasible set.

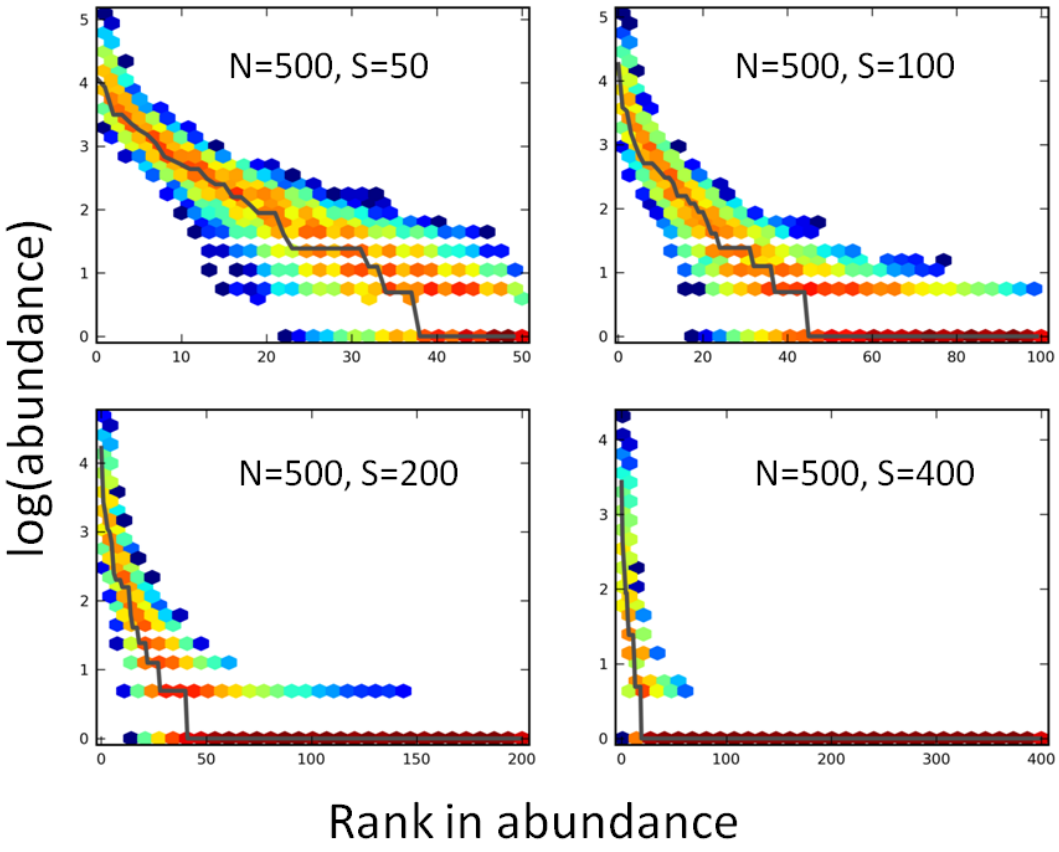


Figure 3. Additional plots, similar to those in Figure 1, reveal the tendency for feasible sets to be dominated by similarly shaped hollow-curve frequency distributions (right-skewed, modal abundance at lowest abundance classes). Each row applies to a single value of total abundance N and three different values of richness S. **Left column**, Subplots reveal the overall shape of the feasible set when all macrostates are plotted as Preston plot frequency distributions. Each macrostate is plotted in a light shade of grey, so that overlapping macrostates generate a light-to-dark grey shaded image of density within the feasible set. **Center column**, kernel density plots reveal the distribution of skewness across the feasible set. Values above 0.0 indicate right skewness. **Right column**, plots of modal abundance reveal that the smallest abundance classes (e.g. singletons, doubletons) are the most common.

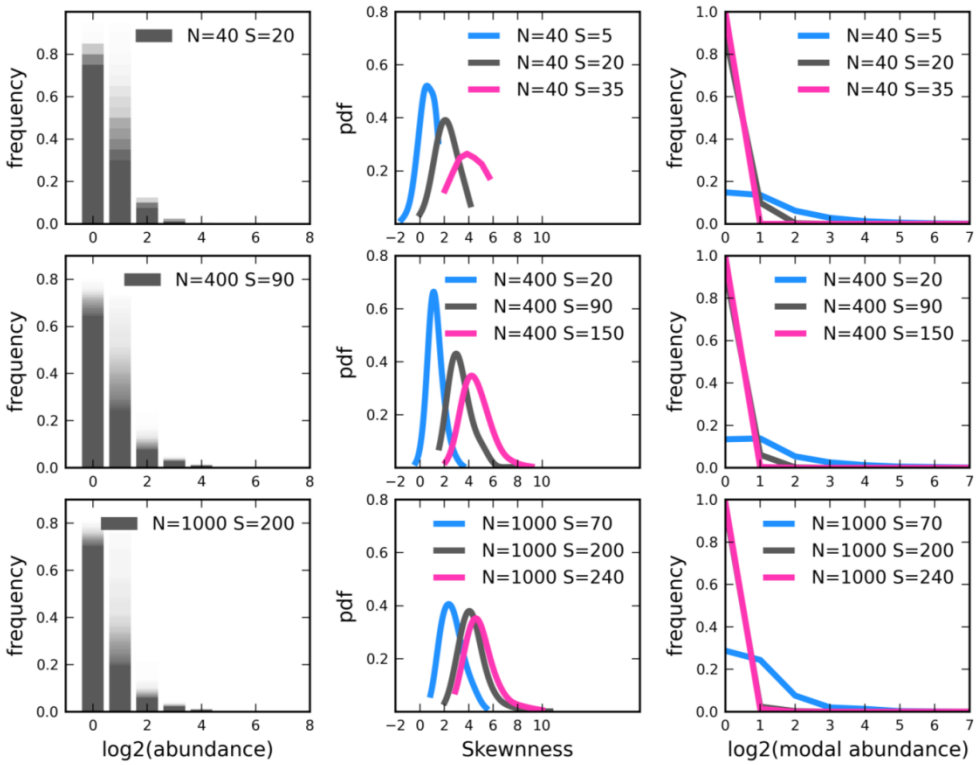


Figure 4. Additional analyses of microbial metagenomic datasets (AQUA, TERA) for species defined at 95 and 99% sequence similarity. These observed vs. expected plots were constructed in the same fashion as Figure 3 in the main manuscript. These plots vary little with percent similarity and reveal the same pattern and similar r-square values to those in Figure 3. A subset of the data from Figure 3 were used to reduce the computational expense of the analysis (which still required 1,380 compute hours). This analysis demonstrates that our results for metagenomes are robust with respect to percent similarity cutoff.

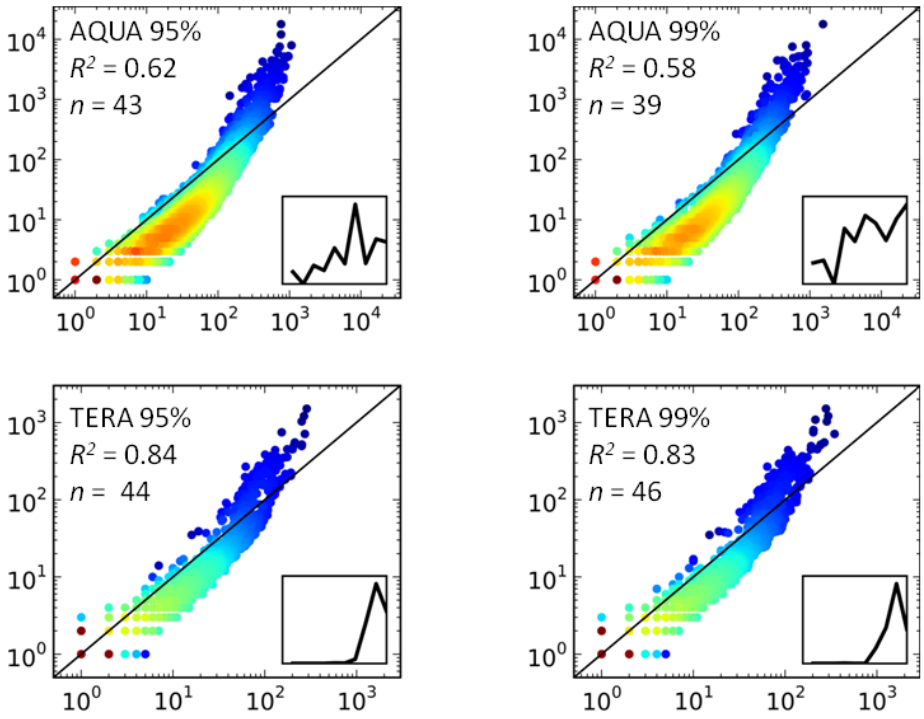
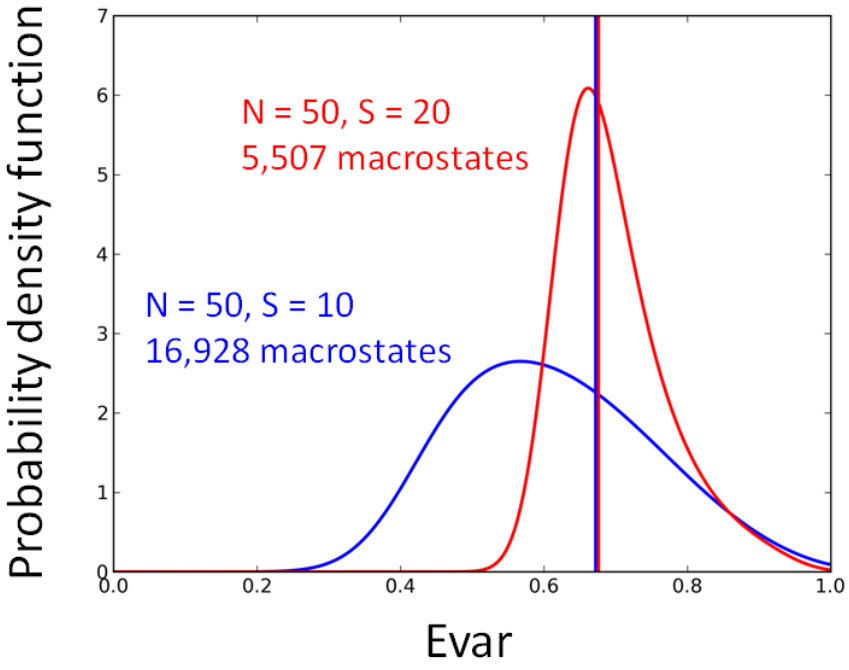


Figure 5. Kernel density plots for values of species evenness (E_{var}) distributed across entire feasible sets. Vertical red and blue lines represent values of E_{var} for two different macrostates in the feasible sets for N = 50 and S = {20, 10}. Despite having similar E_{var}, a value of ~ 0.7 indicating high evenness, the macrostate from the feasible set for N = 50 and S = 10 is exceptionally even with respect to its feasible set, while the macrostate for N = 50 and S = 20 is neither exceptionally even or uneven, and is instead near the central tendency of its feasible set.



References

Amend, A.S., Seifert, K.A., Samson, R., & Bruns, T.D. (2010). Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *P. Natl. Acad. Sci. USA*., 107, 13748-13753.

Chu, H., Fierer, N., Lauber, C.L., Caporaso, J.G., Knight, R. & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol*., 12,2998–3006.

Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A. & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J*., 6, 1007–17.

Flores, G.E., Campbell, J., Kirshtein, J., Meneghin, J., Podar, M., Steinberg, J.I. *et al*. (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol*., 13, 2158-2171.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. *et al*. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386.

National Audubon Society. (2002). The Christmas Bird Count historical results. Retrieved from http://www.audubon.org/bird/cbc.

North American Butterfly Association. (2009). *NABA Butterfly Counts: 2009 Report*, http://www.naba.org.

Sauer, J.R., Hines, J.E., Fallon, J.E., Parkieck, D.J., Ziolkowski, D.J. Jr. & Link, W.A. (2011). *The North American Breeding Bird Survey 1966-2009*. Version 3.23.2011. USGS Patuxent Wildlife Research Center, Laurel, MD.

Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and abundance of mammalian communities. *Ecology*, 92, 2316-2316.

U.S. Department of Agriculture, F.S. (2010). Forest inventory and analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis.

White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

APPENDIX D:

CHAPTER 3: EXPLANATION OF PARTITION FUNCTIONS AND ADDITIONAL FIGURES

The partition function:

Here, the mathematical convention of representing the total as n (instead of q) is followed. The number of partitions of an integer n, i.e. $p(n)$, equals the number of partitions of n having n or less as the largest element (Andrews & Eriksson 2006, Bóna 2006). Likewise, the number of partitions of an integer n into k or less parts, i.e. $p_k(n)$ equals the number of partitions of an integer n into k or less parts and partitions of n having k or less as the largest part (Bóna 2006). Consequently, knowing $p_k(n)$ for $k = \{1, \ldots, n\}$ leads to $p(n)$, the generating function for which is given by:

The recurrence relation for the number of partitions of n into parts having k or less as the largest part (or having k or less parts or k or less as the first part):

Interpretation: Partitions of n into at most k parts either have exactly k parts or they have fewer than k parts. By convention $p(n = 0) = 1$. Also $p_{k=0}(n) = 0$ if $n > 0$.

Simple recursive function showing how the recurrence relation is implemented. This function is too slow for ecologically realistic values of :

```
def parts(n, k):   # find the number of partitions for q with k as the largest part
    if k == 0: return 0
```

if n == 0: return 1

if n < 0: return 0

return parts(n, k - 1) + parts(n - k, k)

Partition functions for Locey and McGlinn (2013) can be found here:

https://github.com/klocey/partitions/blob/master/partitions.py

Explanation of the 'multiplicity' approach, specifically the statement:

The set of partitions of q having a number of k's equal to m contains the set of partitions of $q - k*m$ having less than k as the first part. Consider the 8 partitions of q = 10 having k = 3 as the first element and note that the possible multiples (m) of 3 are 3, 2, 1:

[3, 3, 3, 1]
[3, 3, 2, 2]
[3, 3, 2, 1, 1]
[3, 3, 1, 1, 1, 1]
[3, 2, 2, 2, 1]
[3, 2, 2, 1, 1, 1]
[3, 2, 1, 1, 1, 1, 1]
[3, 1, 1, 1, 1, 1, 1, 1]

According to the above statement, the partitions of q = 10 having three 3's will contain the set of partitions of $q - k*m$ = 10 – 9 = 1 having less than 3 as the first part. Indeed, there is only one possible partition, i.e. [1]. Likewise, the partitions of q = 10 having two 3's will contain the set of partitions of $q - k*m$ = 10 – 6 = 4 having less than 3 as the first part. Indeed, the possible partitions are [2, 2], [2, 1, 1] and [1, 1, 1, 1], i.e. the feasible set for q = 4 having less than three as the largest part. Finally, the partitions of q = 10 having one 3 will contain the set of partitions of $q - k*m$ = 10 – 3 = 7 having less than 3 as the first part. Indeed, the possible partitions are [2, 2, 2, 1], [2, 2, 1, 1, 1], [2, 1,

1, 1, 1, 1] and [1, 1, 1, 1, 1, 1, 1], i.e. the feasible set for q = 7 having less than three as the largest part.

Figure 1. Kernel density curves for **skewness** are derived from random samples of 500 partitions for different combinations of Q and N. These reveal that the integer partitioning algorithms, i.e. Multiplicity, Top-down, Divide and conquer, Bottom-up produce random samples of feasible sets (FS) both when excluding and including parts having zero values, left and right, respectively. We used Sage to generate the entire feasible set (thick red line) for Q = 50 and N = 10. We used the random partitioning function in Sage to generate 500 partitions for Q = 500 and N = 10, which is too large to enumerate in full.

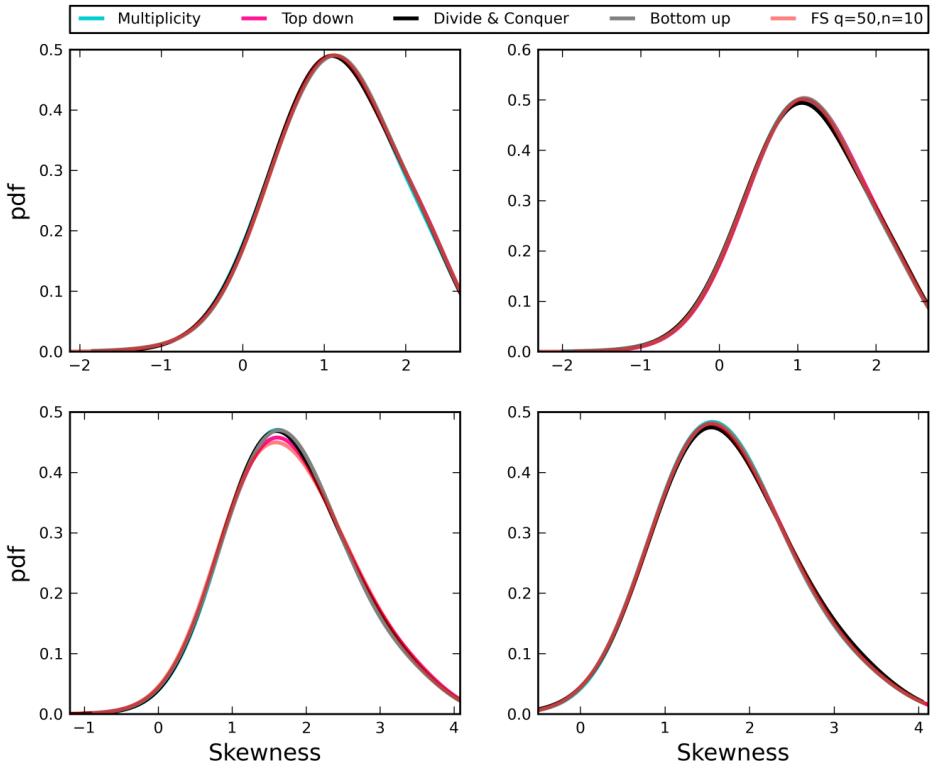


Figure 2. Kernel density curves for the **variance** are derived from random samples of 500 partitions for different combinations of Q and N. These reveal that the integer partitioning algorithms, i.e. Multiplicity (Multi), Top down (T-D), Divide and conquer (D-Q), Bottom up (B-U) produce random samples of feasible sets (F-S) both when excluding and including parts having zero values, left and right, respectively. We used Sage to generate the entire feasible set (thick grey line) for Q = 50 and N = 10. We used the random partitioning function in Sage to generate 500 partitions for Q = 500 and N = 10, which is too large to enumerate in full.

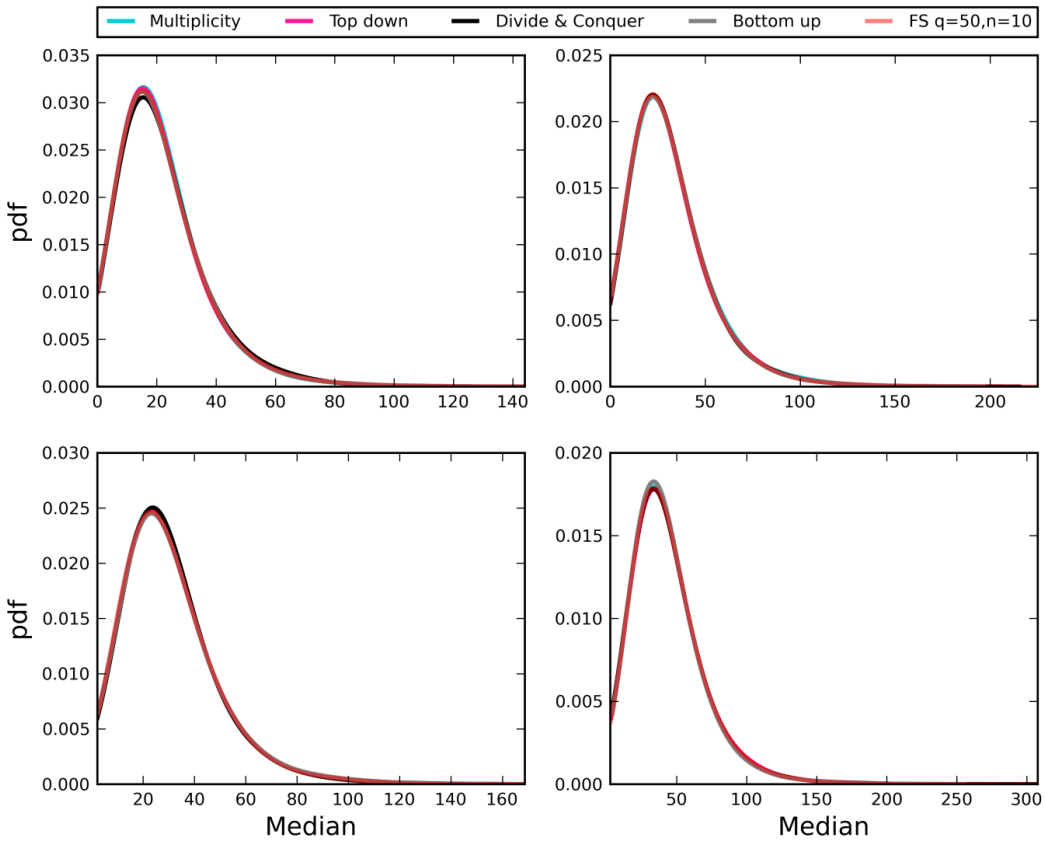
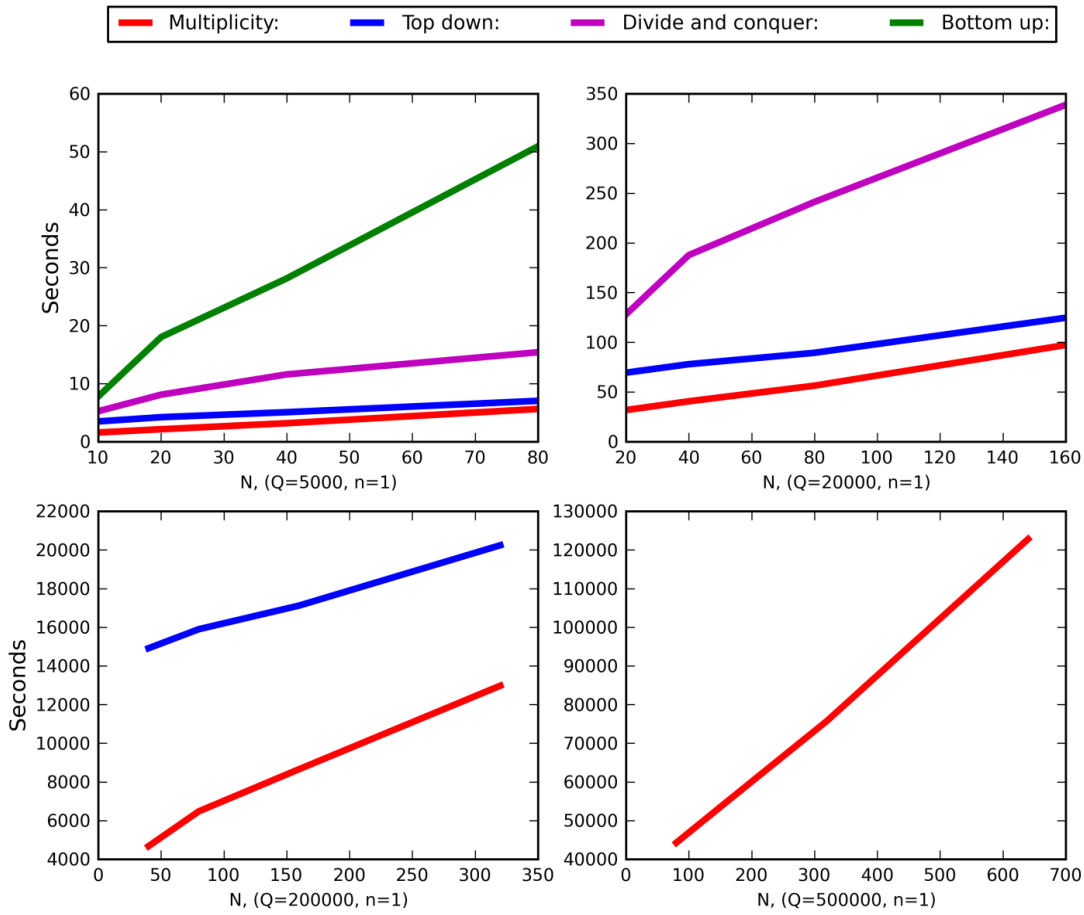


Fig 3. Time required to generate a single random partition (no zero values) for ecologically large combinations of Q and N using the random partitioning algorithms

derived here, implemented in Python. The multiplicity algorithm greatly outperforms the bottom-up and divide-and-conquer methods for these types of Q-N combinations.



CURRICULUM VITAE

KENNETH J. LOCEY
Department of Biology
Utah State University, Logan UT 84322
email: ken@weecology.org
Website: http://kenlocey.weecology.org/

PROFESSIONAL PREPARATION

Institution	Major/Area	Degree and Year
University of Central Oklahoma	Biology	B.S., 2008
Utah State University	Biology	Ph.D., Dec 2013

FELLOWSHIP AND AWARDS

Utah State University, Joseph E. Greaves Scholarship in Biology ($10,000) **2013**

Utah State University School of Graduate Studies, Dissertation Fellowship ($5,000) **2013**

Utah State University, James and Patty MacMahon Scholarship ($500) **2010**

Utah State University, College of Mathematics and Science, George C. Eccles Graduate Fellowship ($66,000) **2008**

University of Central Oklahoma, Lothar Hornuff Outstanding Field Biologist Undergraduate Award ($100) **2008**

PUBLICATIONS

Locey KJ, McGlinn DJ. (**2013**) Efficient algorithms for sampling feasible sets of macroecological patterns. *PeerJ PrePrints* 1:e78v1http://dx.doi.org/10.7287/peerj.preprints.78v1

Locey, K.J., White, E.P. (**2013**) How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, 16: 1177-1185.

White E.P., Baldridge E., Brym Z.T., Locey K.J., McGlinn D.J., and Supp S.R. (**2013**) Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution*, 6, 1-10.

Locey K.J., White E.P. (**2011**) Simple Structural Differences between Coding and Noncoding DNA. *PLoS ONE* 6(2): e14651.

Locey, K.J. (**2010**) Synthesizing traditional biogeography with microbial ecology: the importance of dormancy. *Journal of Biogeography*, 37: 1835–1841.

Stone, P.A., Stone, M.E.B., Stanila, B.D, Locey, K.J. (**2011**) Terrestrial flight response: A new context for terrestrial activity in Sonoran Mud Turtles. *The American Midland Naturalist* 165(1):128 -136. 2011

Stabler, B.L., Johnson, W.L., Locey, K.J., and Stone, P.A. (**2011**) A comparison of Mediterranean Gecko (*Hemidactylus turcicus*) populations in two temperate zone urban habitats. *Urban Ecosystems*. 15:653-666.

Stanila, B.D., Locey, K.J. (**2009**) *Sceloporus jarrovii* (Yarrow's Spiny Lizard). Aquatic escape behavior. *Herpetological Review.* 40 (2): 226.

Locey, K.J., Stone, P.A. (**2008**) Ontogenetic factors affecting diffusion dispersal in the introduced Mediterranean Gecko, *Hemidactylus turcicus* . *Journal of Herpetology* 42:593 -599.

Locey, K.J., Butler C.J., Martin, D.L. (**2008**) *Ctenosaura pectinata* (Spiny-tailed Iguana) population status. *Herpetological Review* 39: 348-349.

Stanila, B.D., Locey, K.J., Stone, P.A. (**2008**) *Kinosternon sonoriense diet Herpetological Review*. 39 (3): 345.

Locey, K.J., Stone, P.A. (**2006**) Factors affecting range expansion in the introduced Mediterranean Gecko, *Hemidactylus turcicus* . *Journal of Herpetology* 40:526 -530.

Locey, K.J., Stone P.A. (**2007**) *Hemidactylus turcicus* (Mediterranean Gecko). *Nesting. Herpetological Review* 38:455–456.

SYNERGISTIC ACTIVITIES

- Guest Speaker, College of Math and Science, University of Central Oklahoma, Title: Introducing Macroecology and Uneven Distributions of Wealth, A Common Pattern in Nature. Nov **2012**
- Invited Participant: USGS Powell Center working group "Next Generation Microbial Indicators of Ecosystem Function" (**2013**)
- Classroom-based online mentor for high school age youths (http://pa-ementor.org/) focusing on post-secondary success, college, and career plans.

GRANTS

PiCloud Academic Research Grant ($500, 10,000 compute hours) **2012**
Amazon Web Services Research Grant (Ethan White and Ken Locey) ($7,800) **2011**
American Museum of Natural History, Theodore Roosevelt Grant ($2,400) **2010**

RECENT PRESENTATIONS

2013, Ecological Society of America, poster
2013, Ecological Society of America, Ignite Session talk