Self-Dissimilarity: An Empirical Measure of Complexity

David H. Wolpert
William G. Macready

SFI WORKING PAPER: 1997-12-087

SANTA FE INSTITUTE

Self-Dissimilarity: An Empirical Measure of Complexity

David H. Wolpert*
NASA Ames Research Center
Caelum Research
MS269-2
Moffett Field, CA, 94035
William G. Macready†
Bios Group LP
317 Paseo de Peralta
Santa Fe, NM, 87501

For systems usually characterized as complex/living/intelligent, the spatio-temporal patterns exhibited on different scales differ markedly from one another. (*E.g.*, the biomass distribution of a human body looks very different depending on the spatial scale at which one examines that biomass.) Conversely, the density patterns at different scales in non-living/simple systems (*e.g.*, gases, mountains, crystal) do not vary significantly from one another.

Such self-dissimilarity can be empirically measured on almost any real-world data set involving spatio-temporal densities, be they mass densities, species densities, or symbol densities. Accordingly, taking a system's (empirically measurable) self-dissimilarity over various scales as a complexity "signature" of the system, we can compare the complexity signatures of wholly different kinds of systems (*e.g.*, systems involving information density in a digital computer *vs.* systems involving species densities in a rainforest, *vs.* capital density in an economy *etc.*). Signatures can also be clustered, to provide an empirically determined taxonomy of kinds of systems that share organizational traits. Many of our candidate self-dissimilarity measures can also be calculated (or at least approximated) for physical models.

The measure of dissimilarity between two scales that we finally choose is the amount of extra information on one of the scales beyond that which exists on the other scale. It is natural to determine this "added information" using a maximum entropy inference of the pattern at the second scale, based on the provided pattern at the first scale. We briefly discuss using our measure with other inference mechanisms (*e.g.*, Kolmogorov complexity-based inference, fractal-dimension preserving inference, etc.).

1 Introduction

Historically, the concepts of life, intelligence, culture, and complexity have resisted all attempts at formal scientific analysis. Indeed, there are not even widely agreed-upon formal definitions of those terms [1, 2]. Why is this?

We argue that the underlying problem is that most of the attempted analyses have constructed an extensive formal model of the underlying processes before considering any particular set of experimental data. Rather than being data-driven, such model-driven approaches are based on insight, or experience gleaned in other fields. For example, some proposed definitions of complexity are founded on statistical mechanical models from physics [3], while others model the underlying processes using computer science abstractions like finite automata [4] or universal Turing machines [5, 6] (see also [7]). None of these model choices arose from consideration of any particular experimental data.

This contrasts with the more empirical approach that characterized the (astonishingly successful) growth of the natural sciences. This approach begins with the specification of readily measurable "attributes of interest" of real-world phenomena, followed by empirical observation of the inter-relationships of those attributes in real-world systems. *Then* there is an attempt to explain those inter-relationships via a theoretical model. For the most part, the natural sciences were born of raw experimental data and a need to explain it, rather than from theoretical musing.

To elaborate this contrast, note that unlike data-first approaches, model-first approaches suffer from being intrinsically dependent upon the (ill-constrained) choice of the model of the underlying process, and of how to map that model onto real-world phenomena. This difficulty is especially acute when one wishes to describe *all* naturally occurring complex systems in terms of the same model class. One would expect that extreme care is necessary in deciding on such a universal model class if descriptions in terms of that class are to be broadly fruitful. Such a choice is made all the more difficult if one tries to make it *a priori*, rather than in response to any experimental data.

Another disadvantage of model-first approaches is that before they can assign a complexity to a system, in many respects they require one to already fully understand that system (to the point that the system is formally encapsulated in terms of one's model). So only once most of the work in analyzing the system has already been done can you investigate that system using these proposed measures of complexity. This is surely a prescription for preventing

*Supported by NASA. Email: dhw@ptolemy.arc.nasa.gov
†Supported by Bios Group, LP. Email: wgm@biosgroup.com

the concept of complexity from being particularly useful. In addition, one would expect this character of model-driven measures of complexity to make it very difficult to apply them to real world systems. This is indeed the case — it is astonishing to note that there have been no studies in which a proposed complexity measure has been evaluated for a broad range of real-world (*not* idealized) data.

Finally, these model-driven approaches are prone to degeneration into theorizing and simulating, in isolation from the real world. The models presented above were not forced upon their proponents, as the only way to explain puzzles in some experimental data. This lack of coupling to experimental data vitiates the most important means by which theoretical models can be compared and modified.

Of course, even in a data-driven approach, there is always some modeling component in one's complexity measure, even if it implicit and slight. The important point is that for a data-driven approach, the measure is designed primarily so that it can be applied to the real world, and with a minimum of assumptions concerning the underlying processes generating that data.

We propose one particular data-driven approach to the study of complex systems. To follow any such approach we must first choose our empirically measurable attribute of interest. That is the subject of this paper. Next, we must measure this attribute of interest for many real-world complex systems. That is work in progress. At this point — and at only at this point — we must use the data resulting from those measurements to guide our construction of potential models of the common processes underlying complex systems. That is future work.

In this paper we consider several potential attributes of interest. These candidate attributes arise from observing that most systems that people intuitively characterize as complex/living/intelligent have the following property: *over different space and time scales, the patterns exhibited by a complex system vary greatly, and in ways that are unexpected given the patterns on the other scales.* Accordingly, a system's self-*dis*similarity is the attribute of interest propose be measured — completely devoid of the context of any formal model at this point.

Implicitly then, our thesis is that variation in a system's spatio-temporal patterns as one changes scales is not just a side-effect of what is really going on in a complex system. Rather it is a crucial reflection of the system's complexity, whatever the underlying mathematical structure controlling that complexity. We propose that it is only after we have measured such self-dissimilar aspects of real-world systems, when we have gone on to construct formal models explaining those data, that we will have models that "get at the heart" of complex systems.

The human body is a familiar example of such self-dissimilarity; as one changes the scale of the spatio-temporal microscope with which one observes the body, the pattern one sees varies tremendously. Other examples from biology are how, as one changes the scale of observation, the internal structures of a biological cell, or of an ecosystem, differ greatly from one another. By measuring patterns in quantities other than the mass distribution (*e.g.*, in information distributions), one also finds that the patterns in economies and other cultural institutions vary enormously with scale. Similarly, presumably as one changes one's scale of observation there are large variations in the charge density patterns inside the human brain.

In contrast, simple systems like crystals and ideal gases may exhibit some variation in pattern over a small range of scales, but invariably when viewed over broad ranges of scales the amount of variation falls away. Similarly, viewed over a broad range of spatio-temporal scales (approximately the scales from complexes of several hundred molecules on up to microns), a mountain, or a chair, exhibits relatively little variation in mass density patterns. As an extreme example, relative to its state when alive, a creature that has died and decomposed exhibits no variation over temporal scales, and such a creature also exhibits far less variation over spatial scales than it did when alive.

There are a number of apparent contrasts between our proposed empirical approach and much previous work on complexity. In particular, fractals have often been characterized as being incredibly complex due to their possessing nontrivial structure at all different scales; in our approach they are instead viewed as relatively simple objects since the structure found at different scales is always the *same* (from an information-theoretic perspective).

Similarly, a cottage industry exists in finding self-similar degrees of freedom in all kinds of real-world systems, some of which can properly be described as complex systems. Our thesis is that independent of such self-similar degrees of freedom, it is the alternative self-dissimilar degrees of freedom which are more directly important for analyzing a system's complexity. We hypothesize that, in large measure, to concentrate on self-similar degrees of freedom of a complex system is to concentrate on the degrees of freedom that can be very compactly encoded, and therefore are not fundamental aspects of that system's complexity.

As an example, consider a successful, flexible, modern corporation, a system that is "self-similar" in certain variables ([9]). Consider such a corporation that specializes in an information processing service of some sort, so that its interaction with its environment can be characterized primarily in terms of such processing rather than in terms of gross physical manipulation of that environment. Now hypothesize that in *all* important regards that corporation is self-similar. Then the behavior of that corporation — and in particular its effective adaptation to and interaction with its environment — is specified using the extremely small amount of information determining the scaling. In such a situation, one could replace that adaptive corporation with a very small computer program based on that scaling information, and the interaction with the environment would be unchanged. The patent absurdity of this claim demonstrates that *what is most important* about a corporation is not captured by those variables that are self-similar.

More generally, even if one could find a system commonly viewed as complex that was clearly self-similar in all important regards, it is hard to see how the same system wouldn't

be considered even more "complex" if it were self-dissimilar. Indeed, it is hard to imagine a system that is highly self-dissimilar in both space and time that wouldn't be considered complex. Self-dissimilarity would appear to be a sufficient condition for a system to be complex, even if it is not a necessary condition.

In Section 2 we further motivate why self-dissimilarity is a good measure of complexity. Section 3 then takes up the challenge of formalizing some of these vague notions. The essence of our approach is the comparison of spatio-temporal structure at different scales. Since we adopt a strongly empirical perspective, how to infer structure on one scale from structure on another is a central issue. This naturally leads to the probabilistic perspective introduced in that section. Next, in Section 4 we discuss computational issues that arise in measuring self-dissimilarity on real-world data sets. Finally, Section 5 presents an overview of future work, and in particular of some of the empirical work we are currently embarking upon measuring real-world self-dissimilarity.

It is worth emphasizing that we wake no claim whatsoever that self-dissimilarity captures all that is important in complex systems. Nor do we even wish to identify self-dissimilarity with complexity. We only suggest that self-dissimilarity is an important component of complexity, one with the novel advantage that it can actually be evaluating for real-world systems.

2 Self-Dissimilarity

In the real world, one analyzes a system by first being provided information (*e.g.*, some experimental data) in one space, and then from that information making inferences about the full system living in a broader space. The essence of our approach is to characterize a system's complexity in terms of how the inferences about that broader space differ from one another as one varies the information-gathering spaces. In other words, our approach is concerned with characterizing how readily the full system can be inferred from incomplete measurements of it. Violent swings in such inferences as one changes what is measured — large self-dissimilarity — constitute complexity for us.

2.1 Why should complex systems be self-dissimilar?

Before turning to formal definitions of self-dissimilarity we speculate on why self-dissimilarity might be an important indicator of complexity. Certainly self-dissimilar systems will be *interesting*, but why should they also coincide with what are commonly considered to be complex systems?

Most systems commonly viewed as complex/interesting have been constructed by an evolutionary process (*e.g.* life, culture, intelligence). If we assume that there is some selective advantage for efficient information processing in such systems, then we are logically led to consider systems which process information in many different ways on many spatio-temporal scales, with those different processes all communicating with one another. Such systems are in a certain sense maximally dense with respect to how much information processing they achieve in a given region. Systems processing information similarly on different scales, or even worse not exploiting different scales at all, are simply inefficient in their information-processing capabilities.

To make maximal use of the different information processes at different scales, there must be efficient communication between those processes. Such inter-scale communication is common in systems usually viewed as complex. For example, typically the effects of large scale occurrences (like broken bones in organisms) propagate to the smallest levels (stimulating bone cell growth) in complex systems. Similarly, slight changes at small scales (the bankruptcy of a firm, or the mutation of a gene) can have marked large-scale (industry-wide, or body-wide) effects in such systems.

Despite the clear potential benefits of multi-scale information processing, constructing a system which engages in such behavior seems to be a formidable challenge. Even specifying the necessary dynamical conditions (*e.g.*, a Hamiltonian) for a system to be able to support multi-scale information processing appears difficult. Here we merely assume that nature has stumbled upon solutions to this problem. Our present goal is only to determine how to recognize and quantify such multi-scale information processing in the first place, and then to measure such processing in real-world systems.

This perspective of communication between scales suggests that there are upper bounds on how self-dissimilar a viable complex system can be. Since the structure at one scale must have meaning at another scale to allow communication between the two, presumably those structures cannot be *too* different. Also, complex systems arising from an evolutionary process must be robust. The effects of random perturbations on a particular scale must be isolated to one or a few scales lest the full system collapse. To this extent scales must be insulated from each other. Accordingly, as a function of the noise inherent in an environment, there may be very precise and constrained ways in which scales can interact in robust systems. If so it would be hoped that when applied to real-world complex systems a self-dissimilarity measure would uncover such a modularity of multi-scale information processing.

This perspective also gives rise to some interesting conjectures concerning the concept of intelligence. It is generally agreed that any "intelligent" organism has a huge amount of extra-genetic information concerning the outside world in its brain. In other words, the information processing in the brain of an intelligent organism is tightly and extensively coupled to the information processing of the outside world. So to an intelligent organism, the outside world — which is physically a scale up from the organism — has the kind of information coupling with the organism that living organisms have within their own bodies.

So what is intelligence? From this perspective, it is a system that is coupled to the broader external world exactly as though it were a subsystem of a living body consisting of that broader world. In other words, it is a system

whose relationship with the outside world is similar to its relationship with its own internal subsystems. An intelligence is a system configured so that the border of what-is-a-living/complex-organism extends beyond it, to the surrounding environment.

2.2 Advantages of this approach

The reliance on self-dissimilarity as a starting point for a science of complexity has advantages beyond being largely data-driven. Puzzles like how to determine whether a system "is alive" are rendered mute under such an approach. We argue that such difficulties arise from trying to squeeze physical phenomena into pre-existing theoretical models (*e.g.*, for models concerning "life" one must identify the atomic units of the system, define what is meant for them to reproduce, *etc.*). Taking an empirical approach though, life is a characteristic signature of a system's self-dissimilarity over a range of spatio-temporal scales. Highly complex living systems exhibit highly detailed, large self-dissimilarity signatures, while less complex, more dead systems exhibit shallower signatures with less fine detail. We argue that life is more than a yes/no bit, and even more than a real number signifying a degree — it is an entire signature. In addition to obviating semantic arguments, adopting this point of view opens new fields of research. For example, one can meaningfully consider questions like how the life-signature of the biosphere changes as one species (*e.g.*, humans) takes over that biosphere.

More generally, self-dissimilarity signatures can be used to compare entirely different kinds of systems (*e.g.*, information densities in human organizations versus mass distributions in galaxies). With this complexity measure we can, in theory at least, meaningfully address questions like the following: How does a modern economy's complexity signature compare to that of the organelles inside a prokaryotic cell? What naturally occurring ecology is most like that of a modern city? Most like that of the charge densities moving across the internet? Can cultures be distinguished according to their self-dissimilarity measure? Can one reliably distinguish between different kinds of text streams, like poetry and prose, in terms of their complexity?

In addition, by concentrating on self-dissimilarity signatures we can compare systems over different regions of scales, thereby investigating how the complexity character itself changes as one varies the scale. This allows us to address questions like: For what range of scales is the associated self-dissimilarity signature of a transportation system most like the signature of the current densities inside a computer? How much is the self-dissimilarity signature of the mass density of the astronomy-scale universe like that of an ideal gas when looked at mesoscopically?

In fact, by applying the statistical technique of clustering to self-dissimilarity signatures, we should be able to create taxonomies ranging over broad classes of real-world systems. For example, self-dissimilarity signatures certainly will separate marine environments (where the density of organisms is similar to the density of the environment) from terrestrial environments (where the densities of organisms is quite different from the density of their environment). One would also suspect that such signatures should divide marine creatures from terrestrial ones, since the bodily processes of marine creatures observe broad commonalities not present in terrestrial creatures (and vice-versa). Certainly one would expect that such signatures could separate prokaryotes from eukaryotes, plants from animals, etc. In short, statistical clustering of self-dissimilarity signatures may provide a purely data-driven (rather than model-driven or — worse still — subjective) means of generating a biological taxonomy. Moreover, we can extend the set of signatures being clustered far beyond biological systems, thereby creating, in theory at least, a taxonomy of all natural phenomena. For example, not only could we cluster cultural institutions. (Do Eastern and Western socio-economic institutions break up into distinct clusters?) We could also cluster the signatures of such institutions together with those of insect colonies. (Do hives fall in the same cluster as human feudal societies, or are they more like democracies?)

Another advantage of the self-dissimilarity concept is that it leads to many interesting conjectures. For example, in the spirit of the Church-Turing thesis, one might posit that any naturally-occurring system with sufficiently complex yet non-random behavior at some scale s must have a relatively large and detailed self-dissimilarity signature at scales finer than s. If this hypothesis holds, then (for example) due to the fact that its large-scale physical behavior (*i.e.*, the dynamics of its intelligent actions) is complex, the human mind *necessarily* has a large and detailed self-dissimilarity signature at scales smaller than that of the brain. Such a scenario suggests that the different dynamical patterns on different scales within the human brain is not some side-effect of how nature happened to solve the intelligence question, given its constraints of noisy carbon-based life. Rather it is fundamental, being required for any (naturally occurring) intelligence. This would in turn suggest that (for example) work on artificial neural nets will have difficulty creating convincing mimics of human beings until those nets are built on several different scales at once.

Of course, one can also compare systems using more traditional characterizations of those systems, like the first several moments. This is an advantage that would accrue to any empirically-driven approach to the study of complex systems. The further advantage of using self-dissimilarity is that it means we are comparing systems based on a quantity intimately connected with the system's information processing and with its complexity.

3 Probabilistic Measures of Self-Dissimilarity

We begin by noting that any physical system is a realization of a stochastic process, and it is the properties of that underlying process that are fundamentally important. This leads us to consider an explicitly probabilistic setting for measuring self-dissimilarity. In particular, the "structure at scale s" is taken to mean the probability distribution over

the various scale s patterns that the process can generate. By incorporating probability theory into its foundations in this way, our approach explicitly reflects the fundamental role that statistical inference (for example of patterns at one scale from patterns at another scale) plays in complexity. In addition, via information theory, it provides us with some very natural candidate measures for the amount of dissimilarity between structures at two different scales (*e.g.*, the Kullback-Leibler [10] distance between those structures).

3.1 Defining the structure at a scale

Assume a nested set of spaces, $\Omega_s \subset \Omega_{s'>s}$. The indices on the spaces are called *scales*. (Such scales are more akin to the widths of the windows with which a system is examined rather than different levels of precision with which it is examined.) For any two scales s_1 and $s_2 > s_1$ we have a set of mappings $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$ labelled by i, each taking elements of Ω_{s_2} to elements of the smaller scale space Ω_{s_1}. Given a probability distribution P_{s_2} over Ω_{s_2} (*i.e.*, a scale s_2 structure) and any single member of the mapping set $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$, we obtain an induced probability distribution over Ω_{s_1} in the usual way. Call that distribution $\rho^{(i)}_{s_1 \leftarrow s_2}(P_{s_2})$, or just $P^{(i)}_{s_1 \leftarrow s_2}$ for short. It is the set of structures at scale s_1 generated by mapping down from the structure at scale s_2.

It is convenient to construct a quantitative synopsis of the set of all of those scale s_1 structures. In particular, if that synopsis is itself a (single) structure, then forming this synopsis puts Ω_{s_1} and Ω_{s_2} on equal footing, in that they are both associated with a single structure. In this paper our synopsis of the $\{P^{(i)}_{s_1 \leftarrow s_2}\}$ will be their weighted average: $\rho_{s_1 \leftarrow s_2}(P_{s_2}) = P_{s_1 \leftarrow s_2} \equiv \sum_i \alpha_{s_1 \leftarrow s_2}(i) P^{(i)}_{s_1 \leftarrow s_2} / \sum_i \alpha_{s_1 \leftarrow s_2}(i)$.[1] $P_{s_1 \leftarrow s_2}$ defines the structure at scale s_1 induced by P_{s_2}.

Presently, we restrict attention to mapping sets such that for any $s_1 < s_2 < s_3$, the set $\{\rho^{(k)}_{s_1 \leftarrow s_3}\}$ is the set of all compositions $\rho^{(i)}_{s_1 \leftarrow s_2} \rho^{(j)}_{s_2 \leftarrow s_3}$. We will call this restriction *composability of mapping sets*. Note that such composability does not quite force $\rho_{s_1 \leftarrow s_3}(P_{s_3})$ to equal $\rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_3}(P_{s_3}))$.[2] In this paper though we focus on mapping sets such that for the scales of interest $P_{s_1 \leftarrow s_3} \approx \rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_3}(P_{s_3}))$. Under this restriction we can, with small error, just write P_{s_1} for any scale of interest s_1. For situations where this restriction holds we will say that we have (approximate) *composability of distributions*.

Example 1: The members of Ω_{s_2} are the sequences of s_2 successive bits. Indicate such a sequence as $\omega_{s_2}(k)$, $1 \leq k \leq s_2$. $\rho^{(i)}_{s_1 \leftarrow s_2}$ is the projective mapping taking any ω_{s_2} to the sequence of s_1 bits ω_{s_1} where $\omega_{s_1}(j) = \omega_{s_2}(j + i)$ for $1 \leq j \leq s_1$, and $1 \leq i \leq s_2 - s_1$. So the $\rho^{(i)}_{s_1 \leftarrow s_2}$ are translations of a simple masking of a subsequence of s_1 bits, with i indicating the translation. With these definitions $P^{(i)}_{s_1 \leftarrow s_2}(\omega_{s_1})$ is the probability that a sequence randomly sampled from Ω_{s_2} will have the subsequence ω_{s_1} starting at its i'th bit. So $P_{s_1 \leftarrow s_2}(\omega_{s_1})$ is the probability that a sequence randomly sampled from Ω_{s_2} will, when sampled starting at a random bit i, have the sequence ω_{s_1}.

Example 2: Again let Ω_{s_2} be the sequence of s_2 successive bits. But now we have non-overlapping mask operators $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$. So $\rho^{(i)}_{s_1 \leftarrow s_2}(\omega_{s_2})$ is the sequence of s_1 consecutive bits $\omega_{s_2}(j + s_1 \times (i-1))$, where $1 \leq j \leq s_1$ and $1 \leq i \leq s_2/s_1$. (It is implicit that s_2 is an integer multiple of s_1.) $P_{s_1 \leftarrow s_2}(\omega_{s_1})$ is now the probability that a sequence randomly sampled from Ω_{s_2} will, when sampled starting at a random bit $s_1 \times (i-1)$, have the sequence ω_{s_1}.

Example 3: This is the same as example 1, except that i nows range up to s_2, with ω_{s_1} for the i's exceeding $s_2 - s_1$ set by periodicity: $\omega_{s_1}(j) = \omega_{s_2}([j + i] \mod [s_2])$.

In example 1, although we have composability of mapping sets, in general we do not have composability of distributions unless s_3/s_2 is quite large. The problem is edge effects arising from the finite extent of Ω_{s_3}. Say $P_{s_3}(\omega_{s_3}) = 1$ for some particular ω_{s_3}; all other elements of Ω_{s_3} are disallowed. Then a subsequence of s_1 bits occurring only once in ω_{s_3} will occur just once in $\{\rho^{(k)}_{s_1 \leftarrow s_3}(\omega_{s_3})\}$, and accordingly is assigned the value $1/(s_3 - s_1)$ by $P_{s_1 \leftarrow s_3}$, regardless of where it occurs in ω_{s_3}. If that subsequence arises at the end of ω_{s_3} and nowhere else it will also occur just once in the set $\{\rho^{(i)}_{s_1 \leftarrow s_2} \rho^{(j)}_{s_2 \leftarrow s_3}(\omega_{s_3})\}$. However if it occurs just once in ω_{s_3}, but away from the ends of ω_{s_3}, it will occur more than once in the set $\{\rho^{(i)}_{s_1 \leftarrow s_2} \rho^{(j)}_{s_2 \leftarrow s_3}(\omega_{s_3})\}$. Accordingly, its value under $\rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_3}(P_{s_3}))$ is dependent on its position in ω_{s_3}, in contrast to its value under $\rho_{s_1 \leftarrow s_3}(P_{s_3})$.

Fortunately, so long as s_3/s_2 is large, we would expect that any sequence of s_1 bits in ω_{s_3} that has a significantly non-zero probability will occur many times in ω_{s_3}, and in particular will occur many times in regions far enough away from the edges of ω_{s_3} so that the edges are effectively invisible. Accordingly, we would expect that the edge effects are negligible under those conditions, and therefore that we have approximate composability of distributions.

For example 2 we have both composability of mapping sets as well as (exact) composability of distributions. Example 3 does not even obey composability of mapping sets.

Note that the mapping from the space of possible P_{s_2} to the space of possible P_{s_1} induced by a particular set of mappings $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$ usually will not be one-to-one. In addition, it need not be onto; there may be P_{s_1}'s that do not live in the space of possible $P_{s_1 \leftarrow s_2}$. In particular, consider example 1 above. Say that $s_1 = 2$. Then $P_{s_1}(\omega_{s_1}) = \delta_{\omega_{s_1},(0,1)}$ is not an allowed $P_{s_1 \leftarrow s_2}$. For such a distribution to exist in the set of possible $P_{s_1 \leftarrow s_2}$ would require that there be sequences ω_{s_2} for which any successive pair of bits is the sequence $(0, 1)$. This results in an immediate contradiction by consideration of possible values the putative ω_{s_2} can have at its bit positions 1 and 2, and then consideration of its

[1] For simplicity, we will usually take a uniform average $\alpha_{s_1 \leftarrow s_2}(i) = 1$.

[2] The problem is that the ratio of the number of times a particular mapping $\rho^{(k^*)}_{s_1 \leftarrow s_3}$ occurs in the set $\{\rho^{(k)}_{s_1 \leftarrow s_3}\}$, divided by the number of times it can be created by compositions $\rho^{(i)}_{s_1 \leftarrow s_2} \rho^{(j)}_{s_2 \leftarrow s_3}$, may not be the same for all k.

possible values for bits 2 and 3.

3.2 Comparison to traditional methods of scaling

There are a number of other ways one might consider defining the structure at a particular scale. In particular, one could imagine modifying any of the several different methods that have been used for studying self-similarity. Although we plan to investigate those methods, it is important to note that they often have aspects that make them appear problematic for the study of self-dissimilarity. For example, one potential approach would start by decomposing the full pattern at the largest scale into a linear combination of patterns over smaller scales, as in wavelet analysis for example ([8]). One could then measure the "mass" of the combining coefficients for each scale, to ascertain how much the various scales contribute to the full pattern. However such an approach has the difficulty that comparing the mass associated with the patterns at a pair of scales in no sense directly compares the patterns at those scales. At best, it reflects — in a non-information-theoretic sense — how much is "left over" and still needs to be explained in the full scale pattern, once one of the smaller scale patterns is taken into account.

Many traditional methods for studying self-similarity rely on scale-indexed blurring functions (*e.g.* convolution functions, or even scaled and translated mother wavelets) B_s that wash out detail at scales finer than s (for example by forming convolutions of the distribution with such blurring functions). With all such approaches one compares some aspect of the pattern one gets after applying B_s to one's underlying distribution, to the pattern one gets after applying $B_{s' \neq s}$. If after appropriate rescaling those patterns are the same for all s and s' then the underlying system is self-similar.

There are certain respects shared by our approach and these alternative approaches. For example, usually a set of spaces $\{\rho^{(i)}_{s_1 \leftarrow s_2} \Omega_{s_2}\}$ are used by those alternative approaches in defining the structure at a particular scale. (Often those spaces are translations of one another, corresponding to translations of the blurring function.)

However, unlike these traditional approaches our approach makes no use of a blurring function. This is important since there are a number of difficulties with using a blurring function to characterize self-dissimilarity. One obvious problem is how to choose the blurring function, a problem that is especially vexing if one wishes to apply the same (or at least closely-related) self-dissimilarity measure to a broad range of systems, including both systems made up of symbols and systems that are numeric. Indeed, for symbolic spaces how even to define blurring functions in general is problematic. This is because the essence of a blurring function B_s is that for any point x, applying B_s reduces the pattern over a neighborhood of width s about x to a single value. There is some form of average or integration involving that blurring function that produces the pattern at the new scale — this is how information on smaller scales than s is washed out. But what general rule should one use to reduce a symbol sequence of width s to a single symbol?

More generally, even for numeric spaces, how should one deal with the statistical artifacts that arise from the fact that the probability distribution of possible values at a point x will differ before and after application of blurring at x? In traditional approaches, for numeric spaces, this issue is addressed by dividing by the variance of the distribution. But that leaves higher order moments unaccounted for, an oversight that can be crucial if one is quantifying how patterns at two different scales differ from one another.

Such artifacts reflect two dangers that should be avoided in a self-dissimilarity measure: i) changes in the underlying statistical process that don't affect how we view the process' self-dissimilarity should not modify the value the self-dissimilarity measure assigns to that process; and ii) changes in the underlying process that modify how we view the self-dissimilarity of the process should alter the value assigned to that process by the candidate measure. In general, unless the measure is derived in a first principles fashion directly from the concept of self-dissimilarity, we can never be sure that the measure is free of such artifacts.

Although in future work we plan to explore the utility of the traditional approaches to defining structure at a scale, our current focus is on the approach outlined above, since it is designed to avoid artifacts as much as possible. In particular, with our approach there is no blurring function, and the problems inherent in such functions are avoided. Intuitively, our approach accomplishes this by having the information at scale s_2 be a superset of the information at any scale $s_1 < s_2$. This is clarified in the following section.

3.3 Comparing structures at different scales

Assume a known distribution P_{s_1} that for example may have been constructed via the operator $\rho_{s_1 \leftarrow s_3}$ from P_{s_3}.[3] Suppose we are interested in the distribution on the scale s_3. Then via Bayes' theorem, our scale s_1 distribution fixes a posterior distribution over the elements of $\omega_{s_3} \in \Omega_{s_3}$:

$$\begin{aligned} P(\omega_{s_3}|P_{s_1}) &= \int dQ \; P(\omega_{s_3}|P_{s_3} = Q) \, P(P_{s_3} = Q|P_{s_1}) \\ &= \int dQ \; Q(\omega_{s_3}) \, P(P_{s_3} = Q|P_{s_1}) \\ &= \frac{\int dQ \; Q(\omega_{s_3}) \, P(P_{s_1}|P_{s_3} = Q) P(P_{s_3} = Q)}{\int dQ \; P(P_{s_1}|P_{s_3} = Q) \, P(P_{s_3} = Q)} \end{aligned}$$

where in the usual Bayesian way $P(P_{s_3} = Q)$ is a prior over the real-valued multidimensional vector P_{s_3}.

So if we know that $P_{s_1} = \rho_{s_1 \leftarrow s_3} P_{s_3}$, then

$$P(\omega_{s_3}|P_{s_1}) = \frac{\int dQ \; Q(\omega_{s_3}) \, \delta\big(P_{s_1} - \rho_{s_1 \leftarrow s_3}(P_{s_3})\big) \, P(P_{s_3} = Q)}{\int dQ \; \delta\big(P_{s_1} - \rho_{s_1 \leftarrow s_3}(P_{s_3})\big) \, P(P_{s_3} = Q)} \tag{1}$$

[3] For a finite space Ω_{s_1} such a distribution is a finite set of real numbers.

In practice, rather than set the prior $P(P_{s_3} = Q)$ and try to evaluate these integrals, one might approximate this fully Bayesian approach, for example via MAXENT [11]), MDL [12], or by minimizing algorithmic complexity [14]. Whatever scheme we use, we will write $\Gamma_{s \to s'}$ to indicate such an *inference mechanism*'s guess for $P_{s'}$, based on a provided structure P_s.

In a similar fashion, a scale s_2 distribution can also be used to infer a posterior estimate of the structure P_{s_3}, where $s_1 < s_2 \leq s_3$. Often s_3 is set in some manner by the problem at hand, and in particular, we can have $s_2 = s_3$. But this is not required by the general formulation.

Once we have calculated both $\Gamma_{s_1 \to s_3}(P_{s_1})$, the scale-$s_1$-inferred structure over Ω_{s_3}, and $\Gamma_{s_2 \to s_3}(P_{s_1})$, the scale-$s_2$-inferred structure over Ω_{s_3}, we have translated both our information concerning Ω_{s_1} and our information concerning Ω_{s_2} into two new sets of information, both of which concern the same space, Ω_{s_3}. At this point we can directly compare the two sets of information concerning scales s_1 and s_2. In this way we can quantify how dissimilar the structures over s_1 and s_2 are, as the amount of information we have from scale s_2 that goes beyond what we have from scale s_1.

So as the next step we must choose a scalar-valued function Δ_{s_3} that measures a distance between probability distributions over Ω_{s_3}. Intuitively, $\Delta_s(Q_s, Q'_s)$ should measure the difference in how much information concerning Ω_s exists in Q and how much exists in Q'. Accordingly Δ_{s_3} should satisfy some simple requirements. For example, it is reasonable to require that for a fixed P_s, $\Delta_s(P_s, Q_s)$ is minimized by setting Q_s to equal P_s. Also, for $P_{s'} = \rho_{s' \leftarrow s}(P_s)$ and $Q_{s'} = \rho_{s' \leftarrow s}(Q_s)$, in some circumstances it might be appropriate to require that $\Delta_s(P_s, Q_s) \geq \Delta_{s'}(P_{s'}, Q_{s'})$.

As an example, $\Delta_s(Q_s, Q'_s)$ might be the magnitude of the difference between Kullback-Leibler distances [10] $D(P_s \| Q_s)$ and $D(P_s \| Q'_s)$ with P_s being the implicit true distribution over Ω_s. (Formally then, we should write $\Delta_s(Q_s, Q_{s'}; P_s)$).) For this choice of Δ and with $s_3 = s_2$, $\Gamma_{s_2 \to s_3}(P_{s_2}) = P_{s_3}$, so $\Delta_{s_3}(\Gamma_{s_2 \to s_3}(P_{s_2}), \Gamma_{s_1 \to s_3}(P_{s_1})) = D(P_{s_3} \| \Gamma_{s_1 \to s_3}(P_{s_1}))$.[4]

The amount of structure at scale s_3 that is deducible from scale s_2 but not from scale s_1, is then the expected value of the distance Δ_{s_3} between the P_{s_3} distribution inferred from P_{s_1} and the P_{s_3} distribution inferred from P_{s_2}. Write that expected distance as

$$I_{s_1, s_2; s_3}(P_{s_1}, P_{s_2}) \equiv$$
$$\int dQ_{s_3} \; \Delta_{s_3}\left(\Gamma_{s_1 \to s_3}(P_{s_1}), \Gamma_{s_2 \to s_3}(P_{s_2}); P_{s_3} = Q_{s_3}\right) \times$$
$$P(P_{s_3} = Q_{s_3} | P_{s_1}, P_{s_2}) \; .$$

In particular, $I_{s_1, s_2; s_2}(P_{s_1}, P_{s_2}) = I_{s_1; s_2}(P_{s_1}, P_{s_2}) = \Delta_{s_2}(P_{s_2}, \Gamma_{s_1 \to s_2}(P_{s_1}))$.

If P_{s_1} and P_{s_2} are both produced by mapping down from P_{s_3}, then using Bayes' theorem,

$$I_{s_1, s_2; s_3}(P_{s_3}) =$$
$$\Big[\int dQ_{s_3} \; \Delta_{s_3}\big(\Gamma_{s_1 \to s_3}(\rho_{s_1 \leftarrow s_3}(P_{s_3})),$$
$$\Gamma_{s_2 \to s_3}(\rho_{s_2 \leftarrow s_3}(P_{s_3})); P_{s_3} = Q_{s_3}\big) \times$$
$$\delta(P_{s_1} - \rho_{s_1 \leftarrow s_3}(Q_{s_3})) \times \delta(P_{s_2} - \rho_{s_2 \leftarrow s_3}(Q_{s_3})) \times$$
$$P(P_{s_3} = Q_{s_3}) \Big]$$
$$/$$
$$\Big[\int dQ_{s_3}$$
$$\delta(P_{s_1} - \rho_{s_1 \leftarrow s_3}(Q_{s_3})) \times \delta(P_{s_2} - \rho_{s_2 \leftarrow s_3}(Q_{s_3})) \times$$
$$P(P_{s_3} = Q_{s_3}) \Big] \; .$$

In particular, if our provided information includes P_{s_3} as well as P_{s_1} and P_{s_2}, then that is reflected in the prior $P(P_{s_3} = Q_{s_3})$, and we just get

$$I_{s_1, s_2; s_3}(P_{s_1}, P_{s_2}) =$$
$$\Delta_{s_3}\left(\Gamma_{s_1 \to s_3}(P_{s_1}), \Gamma_{s_2 \to s_3}(P_{s_2}); P_{s_3}\right) \; .$$

$I_{s_1, s_2; s_3}$ is a quantification of how dissimilar the structures at scales s_1 and s_2 are. The dissimilarity signature of a system is the upper-triangular matrix $\Delta_{s_1, s_2} = I_{s_1, s_2; s_3}(P_{s_1}, P_{s_2})$. Large matrix elements correspond to unanticipated new structure between scales.

In addition to restrictions on the distance measure, there are a number of restrictions we might impose on our inference mechanism. For example, it is reasonable to expect that for scales $i < j < k$ that $I_{i,k} \geq I_{i,j}$. Plugging in Equation (1) with $\rho_{i \leftarrow k}$ set equal to $\rho_{i \leftarrow j} \rho_{j \leftarrow k}$ translates this inequality into a restriction on allowed inference mechanisms $P(P_k | P_i)$ and $P(P_k | P_j)$.

3.4 Features of our Measure

Although we are primarily interested in cases where the indices s do indeed correspond to physical scales and the Ω_s to versions of physical spaces mapped down to physical scales, our formalism does not require this, especially if one allows for non-composable mapping sets. Rather our formalism simply acknowledges that in the real world information is gathered in one space, and from that information inferences are made about the full system. The essence of our approach is to characterize a system's complexity in terms of how the inferences about that broader space differ from one another as one varies the information-gathering spaces. In particular, when $s_2 > s_1$, we are measuring the information on scale s_2 necessary for "knitting together" the scale-s_1 patterns.

Accordingly, there are three elements involved in specifying $I_{s_1, s_2; s_3}(P_{s_1}, P_{s_2})$:

1. A set of mapping sets $\{\rho_{s \leftarrow s'; i}\}$ relating various scales s and s', to define what we mean by "structure" at particular scales;

2. A measure of how alike two structures in the same scale are;

[4] Another natural choice for $\Delta_s(Q_s, Q'_s)$ is $D(Q_s \| Q'_s)$. However this could be misleading if neither Q_s nor Q'_s is well-aligned with the true P_s; in such a case the "extra information" is completely spurious. Note though that for $s_3 = s_2$, we again recover $D(P_{s_3} \| \Gamma_{s_1 \to s_3}(P_{s_1}))$ for this alternative definition of Δ.

3. An inference mechanism to estimate structure on one scale based on the structure on another scale.

The choice of these elements can often be made in an axiomatic manner. First, the measure in (2) can often be uniquely determined based on information theory and the kinds of issues that one wishes to investigate. Assuming one has a prior probability distribution over the set of possible states of the system, then for any provided mapping set, one can combine that prior with the measure of (2) to fix the unique Bayes-optimal inference mechanism: The optimal inference mechanism is the one that produces the minimal expected value of the measure in (2) given the information provided by application of the mapping set. For $s_2 = s_3$, $I_{s_1,s_2;s_3} = \Delta_{s_2}(P_{s_2}, \Gamma_{s1 \to s2}(P_{s_1}))$, and for example for the Kullback-Leibler Δ, the Bayes-optimal $\Gamma_{s_1 \to s_2}(P_{s_1})$ is $P(\omega_{s_2} \mid P_{s_1})$, as in Eq. 1. (This is true for many natural choices of Δ; see the discussion on scoring and density estimation in ([13]).)

Finally, given the mapping-set-indexed Bayes-optimal inference mechanisms, and given the measure of (2), one can axiomatically choose the mapping set itself: The optimal mapping set of size K from Ω_s to $\Omega_{s' \neq q}$ is the set of K mappings that *minimizes* the expected value of the self-dissimilarity of the system. In other words, one can choose the mapping set so that the expected result of applying it to a particular Ω_s results in a distribution over $\Omega_{s'}$ that is maximally informative concerning the distribution over Ω_s, in the sense of inducing a small expected value of the measure in (2). At this point all three components of I are specified. The only input from the researcher was what issues they wish to investigate concerning the system, and their prior knowledge concerning the system.

In practice, one might not wish to pursue such a full axiomatization of the choices of (1,2,3). We view the ease with which our measure allows one to slot in portions of such an alternative non-axiomatic approach to be one of the measure's strengths. For example, one could fix (1) and (2), perhaps in some relatively simple manner without much concern for *a priori* justifiability, and then choose the inference mechanism in a more axiomatic manner. This would allow us to incorporate our prior knowledge concerning the system directly into our analysis of its complexity without following the fully axiomatic approach. For example, if we know that the system has certain symmetries (e.g., translational invariance), then those symmetries can be made part of the inference mechanism.

Another advantage of allowing various inference mechanisms is that it allows us to create more refined versions of some of the traditional measures of complexity. For example, consider a real-world scheme for estimating the algorithmic information complexity of a particular observed physical system. Such a scheme would involve gathering a finite amount of data about the system, and then finding small Turing machines that can account for that data [14]. The appropriately weighted distribution of the full patterns these Turing machines would produce if allowed to run forever constitutes an inference for the full underlying system. Self-dissimilarity then measures how the inference for the full system based on minimizing algorithmic complexity subject to observed data varies as one gathers data in more and more refined spaces. Systems with small algorithmic complexity should be quite self-similar according to such a measure, since once a certain quality of data has been gathered, refining the data further (*i.e.*, increasing the window size) will not affect the set of minimal Turing machines that could have produced that data. Accordingly, such refining will not significantly affect the inference for the full underlying system, and therefore will result in low dissimilarity values. Conversely, algorithmically complex systems should posses large amounts of self-dissimilarity. Note also that rather than characterize a system with just a single number, as the traditional use of algorithmic complexity does, this proposed variant yields a far more nuanced signature (the set $\{I_{s_i,s_j}\}$).

Indeed, by appropriate choice of the inference mechanism, our self-dissimilarity measure can be made to closely approximate traditional, blurring-function-based measures of similarity. All that is needed is for the inference mechanism to work by estimating the fractal character of the pattern at scale s_1, and then extrapolate that character upward to scales $s_2 > s_1$.

3.5 Alternative Ways to Exploit Mapping Sets

There are a number of slight variants of the scheme outlined above which we intend to investigate in the future. These variants all have small disadvantages compared to the scheme outlined above, which makes them less appealing, formally at least.

One example of such a variant is to dispense with inference mechanisms, and define

$$I_{s_1,s_2;s_3}(P_{s_1}, P_{s_2}) = \int dQ_{s_3} dQ'_{s_3} P(P_{s_3} = Q_{s_3} \mid \rho_{s_1 \leftarrow s_3}(P_{s_3})) \times P(P_{s_3} = Q_{s_3} \mid \rho_{s_2 \leftarrow s_3}(P_{s_3})) \times \Delta_{s_3}(Q_{s_3}, Q'_{s_3}; P_{s_3}) \; .$$

(N.b., for the Kullback-Leibler Δ and $s_3 = s_2$, this I gives the same value as the one advocated in this paper.) Or alternatively, one might define things in terms of absolute values:

$$I_{s_1,s_2;s_3}(P_{s_1}, P_{s_2}) = |P(P_{s_3} \mid \rho_{s_1 \leftarrow s_3}(P_{s_3}) - P(P_{s_3} \mid \rho_{s_2 \leftarrow s_3}(P_{s_3})|; ,$$

where the two probability distributions are defined as the associated posteriors over P_{s_3}, evaluated at the actual P_{s_3}.

To determine which precise formulation of I to use, note that what we want to know is "the extra information *concerning patterns* ω_{s_3} that is contained in P_{s_2} but not in P_{s_1}". So we are led to consider a communication channel carrying patterns $\omega_{s_3} \in \Omega_{s_3}$ according to P_{s_3}. We must measure the average surprise at those patterns of an observer 2, who only has access to P_{s_2}, and contrast that with the average surprise at the patterns of an observer 1, who only has access

to P_{s_1}. Accordingly, we are concerned with $P(\omega_{s_3} \mid P_{s_2})$ and $P(\omega_{s_3} \mid P_{s_1})$. If in contrast our communication channel were carrying distributions over Ω_{s_3}, and the observers' surprises at those distributions were the quantity of interest, then we would instead be concerned with $P(P_{s_3} \mid P_{s_2})$ and $P(P_{s_3} \mid P_{s_1})$.

4 Computational Issues

In this section we present a brief overview of some of the issues that arise when measuring self-dissimilarity.

4.1 Data-based Inference

As a practical matter, we are rarely given distributions, but only finite data sets. This means that in addition to the inference mechanism relating structures on different scales we must also have a mechanism of inferring structure from finite data. More formally, if $\mathcal{D}_{s_1}$ and $\mathcal{D}_{s_2}$ are the data observed at scales s_1 and s_2, our task is to evaluate $E(I_{s_1,s_2;s_3}(P_{s_1}, P_{s_2})|\mathcal{D}_{s_1}, \mathcal{D}_{s_2}) = \int dP_{s_1} dP_{s_2}$ Prob$(P_{s_1}, P_{s_2}|\mathcal{D}_{s_1}, \mathcal{D}_{s_2})\, I_{s_1,s_2;s_3}(P_{s_1}, P_{s_2})$. This calculation should take into account how the data sets are generated from the underlying distributions. In particular, say $\mathcal{D}_{s_2}$ consists of K sequences $\{\omega_{s_2}^{(k)}\}$, and let N be the number of operators $\rho_{s_1 \leftarrow s_2}^{(i)}$. Then the operators $\rho_{s_1 \leftarrow s_2}^{(i)}$ can be used to map $\mathcal{D}_{s_2}$ down to a data set $\mathcal{D}_{s_1}$ of NK sequences in Ω_{s_1}. If $\mathcal{D}_{s_1}$ is indeed generated from $\mathcal{D}_{s_2}$ this way, rather than by directly sampling the underlying distributions, then the calculation of Prob$(P_{s_1}, P_{s_2}|D_{s_1}, D_{s_2})$ must reflect that. A detailed investigation of this issue is beyond the scope of this paper.

4.2 The Importance of Window Overlap

The inference mechanism used between scales should be synchronized with the choice of mapping sets. For example, for the mapping sets of Example 2 above a reasonable inference mechanism is to treat each successive non-overlapping window at scale s_1 as an independent sample of the underlying P_{s_3}. This leads to $\mathrm{P}(P_{s_3}|P_{s_1})$ expressed as a delta function about a distribution that consists of s_3/s_1 successive products (one for each non-overlapping window) of P_{s_1}: $P(P_{s_3} = Q_{s_3}|P_{s_1}) = \delta\big(Q_{s_3}, \prod_{i=1}^{s_3/s_1} P_{s_1 \leftarrow s_3}^{(i)}\big)$. If one defines $\Delta_{s_3}(P_{s_3}, Q_{s_3})$ to be the absolute value of the difference between the Shannon entropies of the two Ω_{s_3} distributions P_{s_3} and Q_{s_3}, then $I_{s_1,s_2;s_3}(P_{s_3})$ is simply the redundancy [15, 10] between P_{s_3} and the s_3/s_1 copies of P_{s_1}: $H(P_{s_3}) - (s_3/s_1)H(P_{s_1})$. This is a particularly simple and straightforward self-dissimilarity measure and can be calculated in closed form for simple physical systems. In addition, it is well-known how to form the Bayes-optimal estimate of such a quantity from any finite set of data [16].

Unfortunately, this measure can exhibit peculiar behavior in certain systems. For example, say only one ω_{s_3} is allowed and that it is perfectly periodic with period $n < s_3$. Then redundancy falls to 0 as s_1 rises to n; when our window exactly matches the period, we only need know the (single possible) pattern over the window to infer the pattern over the full space Ω_{s_3}. However as s_1 increases further, we again have multiple possible patterns in the width s_1 window. So our redundancy rises back up, before eventually falling to 0 again when s_1 is an integer multiple of n.[5]

In contrast, if we used the mapping sets of Example 1 in which the scale s_1 windows *do* overlap, then for all $s_1 \geq n$ P_{s_1} would force us to conclude that P_{s_3} is a delta function about that one possible ω_{s_3}. Because of this dependence between the distributions over the different windows, we must use a different inference mechanism from the independence-based one. (This raises a number of interesting computational issues, addressed below.)

Ultimately, this pernicious behavior of the mapping set of Example 2 reflects the fact that our width s_1 windows have no overlap, so we are limited in how we can infer the pattern at scale s_3 from that at scale s_1. In essence, the problem is that the mapping sets don't help us choose any way to relate the s_1 distributions occupying different windows. (Which is exactly why we can assume those distributions are independent in our inference mechanism.)

For these reasons mapping sets like those in Example 2 are not suited to our purposes. However, they do have their uses — for example in time-series analysis. As another example, one might wish to identify regions in a space that have structures that are most unlike each other, and then identify sub-regions within those regions that are in turn most unlike each other (though never comparing sub-regions to the original region). Work on this problem has led researchers to consider "heterogeneity", which involves mapping sets $\rho_{s' \leftarrow s}$ that have non-overlapping windows [17].

4.3 The Importance of Using an Inference Mechanism

It is important to realize that simply using overlapping windows does not, by itself, ensure that one has a reasonable self-dissimilarity measure. One must also use an appropriate associated Δ and inference mechanism. For example, consider using overlapping windows, but with $\Delta(P_{s_1}, P_{s_2})$ the weighted differences in the entropies measured directly at the two scales, so there is no inference mechanism being used.

Since we don't use an inference mechanism, we might be concerned that by comparing the two entropy values we're really comparing apples and oranges. One might expect there to be far too many statistical artifacts to correct than can be addressed using some combination weights. More generally, there is no reflection in this measure of the fact that the two distributions being compared concern the same underlying system — that fundamental fact is completely ignored.

Not surprisingly, these difficulties prove fatal, as the following simple argument shows:

[5] As an aside, it is worth noting that interesting variations of this issue involving broken symmetry arise when we require that ω_{s_3} be periodic with period n, but other than that make no restrictions, so that more than one ω_{s_3} is allowed.

1. Let an alphabet contain N symbols, $\{t_1, \ldots, t_N\}$. If the scale s_1 distribution only allows one pattern (*e.g.* only the pattern t_1 is allowed at scale $s_1 = 1$), it completely specifies the scale $s_2 > s_1$ distribution to only allow a single pattern. By any reasonable measure, there is no information at scale s_2 beyond that at scale s_1. The entropies of the two distributions are both 0, so the weighted difference is also 0 (independent of the weights). This establishes that to measure information at scale s_2 beyond that at scale s_1, we are just interested in that weighted difference, and not that weighted difference minus some overall (potentially non-zero) offset.

2. If all patterns at scale s_1 is equally likely the scale s_1 entropy is $s_1 \times ln(N)$. If the scale s_2 distribution also allows all patterns with equal probability, its entropy is $s_2 \times ln(N)$. Since the extra information can reasonably be demanded to be 0 for this case, we see that the weight for a scale must be the reciprocal of the scale.

3. Now assume the pattern at scale s_1 consists of s_1 sequences $\{t_1, t_2, \ldots, t_{s_1}\}, \{t_2, t_3, \ldots, t_{s_1}, t_1\}, \ldots, \{t_{s_1}, t_1, t_2, \ldots, t_{s_1-1}\}$ all with equal probability $(1/s_1)$, with no other sequences allowed. Then the entropy equals $ln(s_1)$. In this case, we also only have s_1 possible sequences at scale s_2 (*e.g.*, one of them is $\{t_1, t_2, \ldots, t_{s_1}, t_1, t_2, \ldots, t_{s_2 mod(s_1)}\}$), since the scale s_1 pattern has uniquely fixed the overall sequence as being a continued repeating of the sequence $\{t_1, t_2, \ldots, t_{s_1}\}$. In this case though the entropies at the two scales are *both* $ln(s_1)$, so the weighted difference between them is $ln(s_1) \times [\frac{1}{s_2} - \frac{1}{s_1}] \neq 0$. Yet there is no extra information at scale s_2 beyond that at scale s_1. So we have a contradiction.

Thus a weighted combination of entropies is insufficient as a measure of "new information".

4.4 Inference with Overlapping Windows

When the windows are allowed to overlap there are some P_{s_1} for which there are *no* compatible P_{s_3}. Fortunately, so long as one is careful in estimating P_{s_1} from the data such a P_{s_1} will not arise, and there are in fact many P_{s_3} compatible with one's estimated P_{s_1}. However this leaves the problem of how to deal with the multiplicity of those compatible P_{s_3}. The proper solution to this problem is given by Equation 1. However evaluating this expression can be highly non-trivial for the large systems commonly of interest (where s_3 can be in the millions). Accordingly, we are led to consider approximations.

One natural approximation is to impose regularization, *i.e.*, impose our prior over the possible P_{s_3} and then calculate the maximum a posteriori (MAP) P_{s_3} conditioned on having the values of multiple sums of the components of P_{s_3} (*i.e.*, the values of the multiple marginalizations of P_{s_3} down to distributions over the multiple windows) all be given by P_{s_1}.[6] An alternative is to incorporate knowledge that P_{s_1} is itself an average over those sums $P^{(i)}_{s_1 \leftarrow s_3}$: perform our maximization subject to only the single constraint that the average of the sums equals P_{s_1}. At that point one could approximate the full integral of interest for measuring self-dissimilarity by taking $\mathrm{P}(P_{s_3}|P_{s_1})$ to be a delta function about that MAP structure.

As an example, if one has a Gaussian $P(P_{s_3})$, then this procedure reduces to least-mean-squares estimation of the $\mathcal{R}^{s_3}$ vector P_{s_3} subject to constraints that a certain $2^{s_1} \times (s_3/s_1)$ of the possible sums of the components of that vector all result in the vector P_{s_1}. (There are 2^{s_1} sums for each window, and s_3/s_1 windows.) Due to finiteness of one's data, only approximate adherence to the data-generated estimate of the constraints is sensible in practice. Accordingly, this procedure reduces to inverting the influence matrix.

A natural alternative is to replace the Gaussian prior with the entropic prior implicit in the MAXENT technique. MAXENT is the natural choice of inference and has been used in this context for tomography. The idea is quite simple: determine P_{s_3} by maximizing the entropy of P_{s_3}, subject to the constraint that integrating out all variables that are in Ω_{s_3} but not in Ω_{s_1} results in the observed P_{s_1}. Let x_{s_3} indicate the variables describing Ω_{s_3} and x_{s_1} describing Ω_{s_1}. Also let $x_{s_3 - s_1}$ indicate the variables in s_3 but not in s_1. Then we consider maximizing the energy functional:

$$E[P_{s_3}, \lambda] = S[P_{s_3}] + \int dx_{s_1} \; \lambda(x_{s_1}) \left[\int dx_{s_3 - s_1} \; P_{s_3}(x_{s_1}, x_{s_3 - s_2}) - P_{s_1}(x_{s_1}) \right].$$

In this expression the λ are a set of Lagrange parameters enforcing the constraint that P_{s_3} integrates down to P_{s_1}, and $S[P_{s_3}]$ is the usual entropy functional.

For the linear bitstrings of the examples above, the MAXENT inference mechanism is equivalent to assuming P_{s_3} is a translation-invariant Markov random field, with symmetric neighborhoods extending $(s_1 - 1)/2$ forward and backward of the central random variable [19]. Combined with calculational techniques like the use of transfer matrices, this connection allows for the exact calculation of the maximum of $E[P_{s_3}, \lambda]$ in certain circumstances. Unfortunately, space limitations do not allow for a discussion of this issue here. A more generally applicable algorithm for approximately maximizing $E[P_{s_3}, \lambda]$ is given in [20].

5 Current Research

Part of our current work consists of theoretical calculations of self-dissimilarity signatures for physical systems (*e.g.*, Ising spin systems). We are also setting up computer code to empirically measure self-dissimilarity signatures for real-world data sets. Many data sets are available upon which to test our measure. Current investigations include: 1) letter, word-type *etc.* distributions in text documents of vari-

[6] The MAP value of a random variable is defined as the mode of the posterior probability distribution over that random variable [18, 11].

ous kinds, including literature, technical articles, postscript, and compressed versions of all of these; 2) brightness density in one-dimensional scans of images (extending our measure to two-dimensional scans is work for the near future); 3) symbol sequences within both coding and non-coding sections of genomes; and 4) internet traffic distributions. In regard to the latter, it is fascinating to note that previous studies have suggested that both intranet and internet traffic is self-similar, meaning that it is dead by our measure [21], 3).

Once these signatures have been determined they can be statistically clustered. In addition to the general kinds of questions mentioned in Section 2, there are many other interesting questions concerning such signature clusters:

1. Will different kinds of text documents (*e.g.*, postscript versus ascii) break up into different clusters? Do different kinds of literature, or documents in different languages, cluster in signature space? How does the signature of compressed text compare to that of uncompressed text? How do signatures change as one examines progressively later portions of text? In particular, which kind of text is more complex? Can some aspect of the type of the underlying text be inferred by looking at the signature of its compressed version? Can this be done for encrypted documents?

2. How do the clusters produced by hierarchical clustering of self-dissimilarity signatures of living organisms compare to the taxonomy modern biology has created?

3. Do coding and non-coding regions in genomes have distinct kinds of signatures? Are there "blips" in those signatures at the scale of genes, at the scale of functional bundles of genes (as in epistasis), etc.?

4. How do signatures based on symbol distributions in genomes relate to signatures based on phenotypic mass distributions? What happens if we extend the analysis to mass distributions inferred from the fossil record? Can clustering over these types of self-dissimilarity signature provide a novel means of phylogenetic tree reconstruction? Do either type of signature exhibit a trend over evolutionary time-scales (*i.e.*, are organisms getting more complex over time)?

In addition to this empirical work there are theoretical issues that require further investigation. For example, what kind of dynamical laws/Hamiltonians are necessary/sufficient for a system to be strongly self-dissimilar? Can a dynamical system be self-dissimilar over one part of its phase space but not another? What if one also varies the interval of scales over which one is examining that phase space? What are the self-dissimilarity signatures of simple physical systems like Ising spins? What is the best way to define self-dissimilarity for spatio-temporal systems (where, for example, all motion is confined to a light-cone, and there are different units for time and space)? How best should one define information processing on a particular scale? How best should one define communication between scales?

Clearly there are many ways in which self-dissimilarity should prove a fruitful concept with which to investigate the natural and artificial world's complex systems.

Acknowledgements We would like to thank Tony Begg, Liane Gabora, Isaac Saias, and Kevin Wheeler for helpful discussions. We also acknowledge support from the Santa Fe Institute during the formative stages of this research.

References

[1] Lloyd, S, "Physical Measures of Complexity", *1989 Lectures in Complex Systems*, (E. Jen ed), Addison-Wesley, 1990.

[2] Casti, J. L., "What if", *New Scientist* **151** (1996), 36–40.

[3] Lloyd, S. and H. Pagels, "Complexity as thermodynamic depth", *Annals of Physics*, (1988), 186–213.

[4] Crutchfield, J. P., "The calculi of emergence", *Physica D*, **75** (1994), 11–54.

[5] Chaitin, G. *Algorithmic Information Theory*, Cambridge University Press, 1987

[6] Solomonoff, R. J., *Inform. Control*, **7**, (1964), 1.

[7] Bennett, C. H. *Found. Phys.*, **16**, (1986), 585.

[8] Holschneider M., *Wavelets an analysis tool*, Oxford (1995).

[9] Stanley One of his Nature papers on companies.

[10] Cover, T. M., and J. A. Thomas, *Elements of information theory*, John Wiley & Sons (1991).

[11] Jaynes E. T., *Probability theory: the logic of science*, fragmentary edition available at ftp://bayes.wustl.edu/pub/Jaynes/book.probability.theory

[12] Buntine, W. "Bayesian back-propagation", *Complex Systems*, **5**, (1991), 603–643.

[13] Bernardo, and Smith, *Bayesian Theory*, John Wiley & Sons (1995).

[14] Schmidhuber, J. "Discovering solutions with low Kolmogorov complexity and high generalization ability", *The Twelfth International Conference on Machine Learning*, (Prieditis and Russel Eds.), Morgan Kauffman, 1995.

[15] Palus, M. "Coarse-grained entropy rates for characterization of complex time series", Santa Fe Institute TR 94-06-040, 1994

[16] Wolpert, D. H., and Wolf, D. R. "Estimating functions of probability distributions from a finite set of samples", *Phys. Rev. E*, **52** (1995), 6841.

[17] Li, W. "The measure of compositional heterogeneity in DNA sequences is related to measures of complexity", to appear in *Complexity*, 1997.

[18] Berger J. O. *Statistical decision theory and Bayesian analysis*, Springer-Verlag (1985)

[19] Kinderman, R. and J. L. Snell, *Markov random fields and their applications*, American Mathematical Society (1980).

[20] Mottershead, C. T. "Maximum Entropy Tomography", *Maximum Entropy and Bayesian Methods*, (Hanson and Silver Eds.), Kluwer Academic, 1996.

[21] Leland W. E., et al. "On the self-similar nature of ethernet traffic", *IEEE/ACM Transactions on Networking*, **2** (1994), 1.