Robust Monomer-Distribution Biosignatures in Evolving Digital Biota

Evan D. Dorn[1] and Christoph Adami[2,3,4]

Abstract

Because organisms synthesize component molecules at rates that reflect those molecules' adaptive utility, we expect a population of biota to leave a distinctive chemical signature on its environment that is anomalous given the local (abiotic) chemistry. We observe the same effect in the distribution of computer instructions used by an evolving population of digital organisms, and we characterize the robustness of the evolved signature with respect to a number of different changes in the system's physics. The observed instruction abundance anomaly has features that are consistent over a large number of evolutionary trials and alterations in system parameters, which makes it a candidate for a non-Earth-centric life diagnostic. Key Words: Artificial life—Digital life— Astrobiology—Evolution—Selection—Biogeochemistry. Astrobiology 11, 959–968.

1. Introduction

WHEN SEARCHING FOR SIGNATURES of extraterrestrial life, one is inevitably drawn into the quandary of extrapolating from terran biochemistry to universal principles (Ward and Benner, 2007). If we stipulate that all forms of life must be chemistry-based and encode information about their world within molecules that evolve according to Darwinian rules, then the set of possible chemistries, while still unlimited, is constrained by universal features (Benner and Switzer, 1999; Bains, 2004; Benner, 2010). In particular, evolved bioorganisms impress a distinctive chemical signature on their environment, because biota synthesize those compounds that are necessary for competition and replication. More generally, biochemical synthesis may be seen as the product of natural selection, as evolution shapes genomes so as to maximize their fitness. As a result, the chemical species that persist in a biotic environment generally deviate from the chemical species we would expect in the absence of life.

Previously, we demonstrated (Dorn *et al.*, 2011) that a quantifiable phenomenon (which we call the Monomer Abundance Distribution Biosignature or MADB) is measurable in both terrestrial biochemicals and in the artificial life system Avida. Patterns of monomer concentration (amino acids or carboxylic acids in the biosphere, or computer instructions in the digital life environment) reliably distinguish between life-bearing and sterile environments.

The biotic patterns appear to reflect evolutionary constraints on the organisms' composition and function, while abiotic patterns reflect thermodynamic and formation-kinetic constraints. The evolutionary constraints are many: in terrestrial biology, for example, proteins must be specific to achieve particular functions, given other proteins and small molecules they interact with. But they must also fold reliably, and if they are soluble in water, they must have a hydrophobic core. Both functional and structural constraints affect which amino acids appear in a sequence (Forsdyke, 2005), and the same is expected for the composition of programs in artificial life.

This study examines the robustness of the MADB and attempts to demonstrate with a higher degree of confidence that it results from selection pressures acting on evolving biota. A robust biosignature is one that manifests reliably in the presence of life. We examine this hypothesis by altering the underlying "thermodynamics" of the artificial life system and measuring the robustness of the MADB as the digital biota evolve under varying conditions. Our goal is not just to demonstrate that the MADB forms but to measure the extent to which its features are the result of selection by quantifying their independence from the underlying physics.

As we will see, the MADB is largely, but not entirely, conserved even as the physics are changed. Examination of which aspects are conserved can be linked to understanding the function, behavior, and adaptive needs of the digital biota.

[1]Computation and Neural Systems, California Institute of Technology, Pasadena, California.
[2]Keck Graduate Institute of Applied Life Sciences, Claremont, California.
[3]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan.
[4]The BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, Michigan.

1.1. The Monomer Abundance Distribution Biosignature (MADB)

The rates of formation and diagenesis of individual chemical monomers (*e.g.,* amino acids, carboxylic acids, or other ligands) in the absence of life are dictated by the laws of formation kinetics and thermodynamics. Consequently, the observed relative abundances of various monomers in an abiotic environment reflect these constraints. For example, when amino acids are formed without life, large and thermodynamically expensive molecules such as valine are always seen at drastically lower concentrations than simpler compounds like glycine and alanine (Miller, 1953, 1957; Kvenvolden *et al.,* 1971; Wolman *et al.,* 1972; McDonald *et al.,* 1994; Munoz-Caro *et al.,* 2002; Dorn *et al.,* 2003).

Organisms, on the other hand, are constrained by their need to reproduce and compete. Biota expend energy to manufacture whatever monomers are necessary to meet a fitness criterion; while synthesizing a particular molecule may be relatively expensive, if it is essential to competition the alternative may be extinction. Therefore, we expect evolved genotypes to synthesize molecules at rates that reflect those molecules' utility in fitness rather than, or in addition to, their thermodynamic cost. This effect has been previously discussed by Lovelock (1965), McKay (2002, 2004), Summons *et al.* (2008), Davies *et al.* (2009), and Shapiro and Schulze-Makuch (2009). McKay (2004) and Davies *et al.* (2009) coined the term "Lego Principle" to describe the specific case that biological systems employ a discontinuous subset of the possible molecules in a family of biochemicals, for example, that terrestrial biota use only two dozen or so out of a much larger number of possible amino acids. Such a signature has also been proposed as a means to discover a "shadow biosphere" on Earth: a form of biochemical life with an independent origin from the life we know today (Davies *et al.,* 2009).

We call any measurable variation between biotic and abiotic monomer concentrations within a chemical family the Monomer Abundance Distribution Biosignature (MADB). The MADB is very pronounced in the terrestrial biosphere and is easily detectable by a number of mathematical techniques (Dorn *et al.,* 2003, 2011).

1.2. Artificial life as a test bed for MADB

One of the fundamental concerns for any putative biosignature is that we have only one biosphere to test it against. Yet an ideal biosignature should be capable of indicating the presence of life regardless of its particular chemistry. Moreover, we cannot conduct experiments to "start evolution over" in the terrestrial biosphere and examine the resulting evolved biochemistry. To overcome these hurdles, we turn to artificial life (in particular *digital life*) for an additional example of life and of evolutionary processes that we can use as a test bed. Because life *in silico* is unrelated to terrestrial biochemistry, this also serves to help abstract our thinking and avoid assumptions resulting from a terrestrial bias.

Here, we use the artificial life platform Avida [an introduction to Avida may be found in a study by Ofria and Wilke (2004); see Adami (2006) for a review of research performed with artificial-life techniques]. Avida organisms ("avidians") are small, self-replicating programs written in a simple programming language; 29 instructions are available in the variant used for this study. The instructions may be seen as analogous to the monomers (such as amino acids) that compose familiar biota because the frequency with which instructions appear within the avidian genomes is a very good proxy for the frequency with which they are executed, that is, the phenotype of the organism. This is due to the linearity of program execution in Avida, where loops are uncommon (except for a single, usually short, loop used for replication). While the instructions/amino acids analogy is inexact, Avida instructions share three key properties with chemical monomers: the "biomass" of avidians is composed of those instructions, and that composition is both inheritable and subject to selection. If the constituents' (whether biochemical or digital) relative concentrations are measurable, these properties are all that is necessary for an MADB to form.

In Avida, instructions are substituted and inserted into genomes via externally imposed mutations, including copy errors, point mutations, and insert and delete mutations. Genomes have multiple options of monomers from which to construct genes, and by default, all instructions appear with equal probability when a mutation is imposed. In this world, this represents a fundamental abiotic process, since avidians cannot affect the mutation rate or the frequency of appearance of any particular instruction. If adaptation did not constrain their abundance, we would expect all 29 instructions to appear in equal proportion in the population.

When the bulk frequency of programming instructions is counted in evolved populations in Avida, we observe a distinct profile that does not reflect either the abiotic parameters of the system or the instruction frequency of the ancestor, which indicates that selection has dictated the monomer abundance pattern of the population. This pattern has largely been consistent over many trials, even though the actual genomes that have evolved may not resemble each other at all. Moreover, the pattern is consistent over a wide array of parameters such as mutation rate and different ancestors. Figure 1 shows the relative distribution of 28 computer instructions in evolved Avida populations that are descendant from two distinct ancestor genotypes [these results are more fully presented by Dorn *et al.* (2011)]. One instruction (NOP-A) is excluded from our analysis; see Methods for a discussion. Note that while the ancestors have very different composition, their descendants have converged to some extent to a common profile, which demonstrates the dominant effect of selection on monomer abundances.

When life is introduced to a formerly abiotic environment, the MADB rapidly overwhelms the preexisting abiotic signature, as seen in Fig. 2. In the experiment that produced Fig. 2, an Avida population was seeded with randomly generated, nonviable genomes and bombarded with a high (lethal) level of point mutations. Single, viable intact organisms were periodically introduced into the environment, while the rate of point mutations was stepped down. When the mutation rate became low enough for organisms to survive, avidians quickly populated the entire landscape, impressing their signature distribution of instructions onto the environment. An initial spike reflects the ratios of instructions present in the ancestor genotype, but this was quickly replaced by an evolved MADB as the organisms adapted.

In this study, we further explored the robustness of the MADB as the fundamental abiotic parameters of the Avida environment were changed. This is important because it could be argued that the MADB observed in terrestrial biochemicals
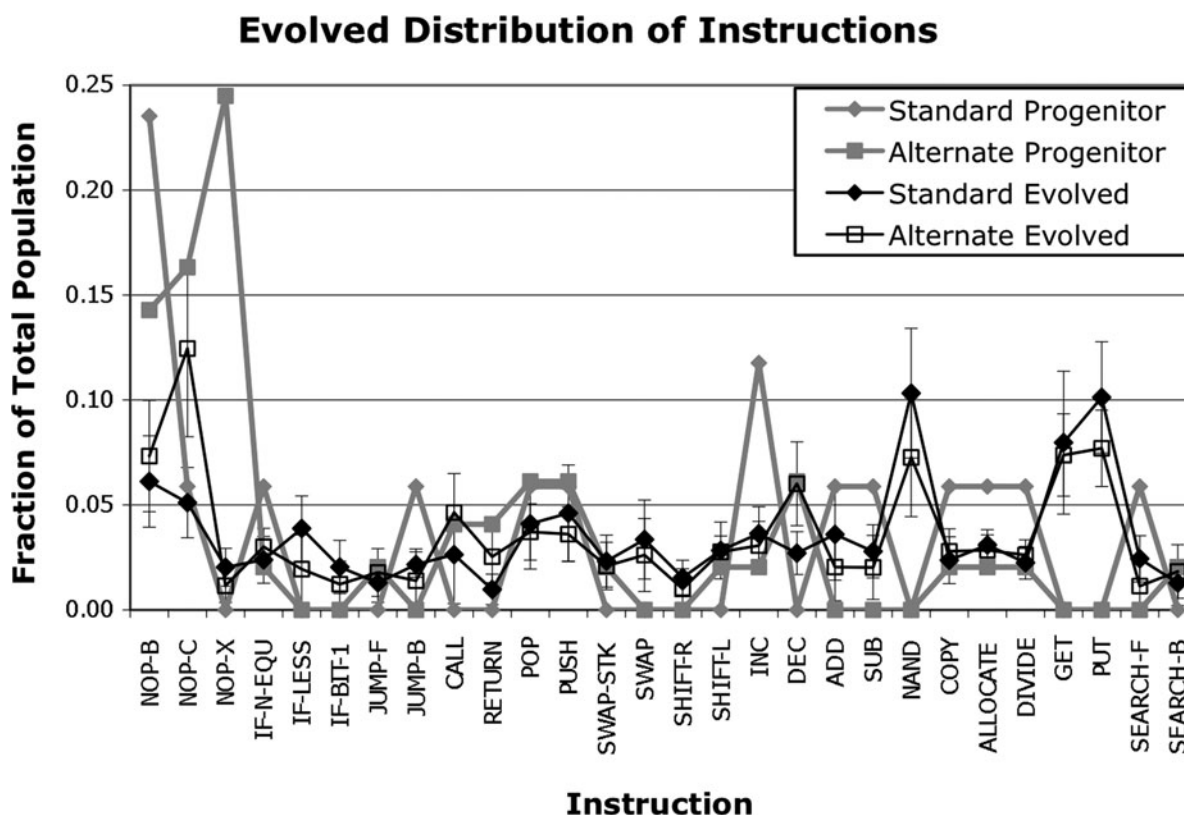
Evolved Distribution of Instructions

FIG. 1. The distribution of instructions in two different ancestor organisms and in populations descended from those ancestors. While the ancestors have very different composition, the composition of the descendants is similar, as the terminal populations have adapted to the same environment. "Evolved" lines represent the average of 25 different evolutionary trials, each sampled after 1500 generations. Error bars are one standard deviation. [Data from the experiment described in (Dorn *et al.*, 2011), in which all instructions were equally available in mutation]. There are no errors for the progenitor distributions, as they are exact.

(*e.g.,* amino acids) is highly dependent on the formation thermodynamics of the individual monomers, and that the pattern would be drastically altered if the costs of synthesis were changed. Alternatively, it is conceivable that the distinction seen between biotic and abiotic patterns is not the product of selection but of some other, unknown function.

In artificial life, we can test these conjectures. Specifically, to study the robustness of the MADB in digital organisms, we alter the availability of each instruction by changing the frequency with which it appears in mutation; this is loosely analogous to altering the formation thermodynamics of amino acids, which thus changes their availability to early life-forms, or to altering the mutational bias on individual nucleotides. If elements of the MADB pattern are retained despite these alterations, it demonstrates that selection is capable of overwhelming the constraints of physics with respect to the composition of organisms in early evolution.

We hypothesize that some instructions' presence (or absence) will be more or less independent of the frequency with which they appear in mutation, which indicates that their appearance frequency in the genome is strongly constrained by a fitness criterion, while other instructions are less strongly constrained. Instructions that convey a strong fitness benefit should be incorporated into genomes rapidly, which would thus ensure that they account for a large proportion of the final population. Anti-adaptive instructions

(*i.e.,* ones that are more often deleterious when appearing as mutations) should be suppressed in the population. We should emphasize that not all deviations from the frequency with which an instruction is created by mutations are adaptive. In any evolutionary unfolding, changes can be due to chance, due to adaptation, or due to historical contingency (Travisano *et al.*, 1995; Joshi *et al.*, 2003; Wagenaar and Adami, 2004). For example, one instruction that could be used as an alternative to another might be "locked in" early during evolution and appear at an increased frequency throughout history when the alternative could just as well have been used. At the same time, neutral drift and chance events could affect instruction frequency, even though such a frequency would not be stable. Because all these effects are expected to shape the frequency distribution of any system undergoing Darwinian evolution, they do not distract from the universality of the MADB.

2. Methods

By default, Avida substitutes new instructions (monomer synthesis) during mutation events with an equal probability for each instruction. However, this is unlikely to be realistic in monomer chemistry, as each monomer can be expected to have its own formation probability. Thus, this "probability of synthesis" should have a bias that reflects the environment's
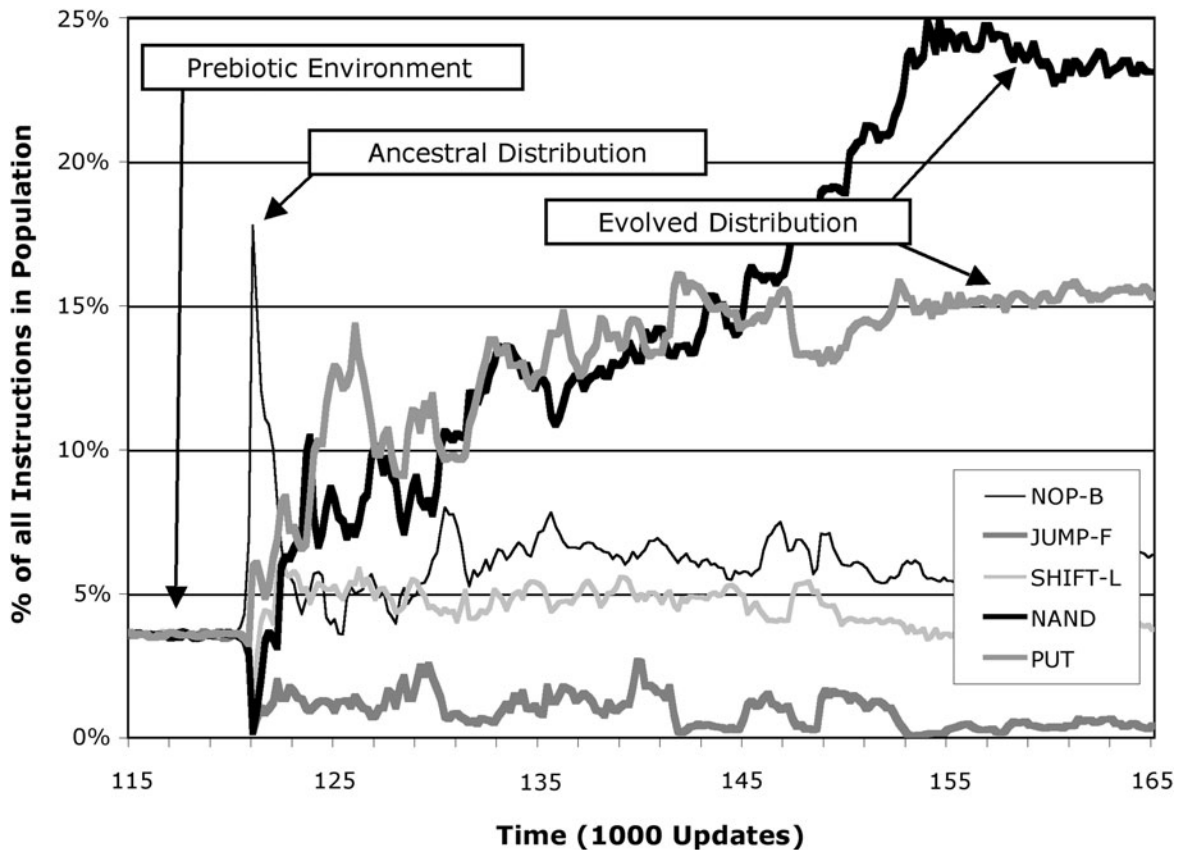
FIG. 2. Evolution of the distribution of six computer instructions as incident self-replicators colonize and adapt to a formerly lifeless environment. At the outset, mutations cause all instructions to be present in roughly equal proportion. The ancestor organism's genome, nearly 20% NOP-B, dominates the early biotic distribution. As the organisms adapt to the environment, a NAND-heavy distribution develops and stabilizes. Instructions such as JUMP-F that are often lethal are strongly suppressed by selection pressure. From Dorn *et al.* (2011).

physics or chemistry. We model different such biases by constructing systematic biases by hand, or else by creating random biases. For this experiment, we created a modified version of Avida version 1.6 that allows the experimenter to specify a probability of substitution for each instruction (in the standard version, each instruction is substituted with equal probability) and that includes a nonstandard output to report the population frequency of each instruction. The code for this version of Avida, along with the configuration files used, are available in the online Supplementary Information (available at www.liebertonline.com/ast) for this article or from the authors.

2.1. Experiments A–D (Figure 3)

In each experiment, a grid of 3600 cells was populated with a 13-instruction simple self-replicating ancestor, whose length can change due to insertion and deletion mutations. This initial population was evolved for 1500 generations, and the bulk frequency of each instruction in the population was quantified every 100 generations. The bulk frequency of an instruction is given by the total number of that instruction in the population divided by the total number of instructions in the population (3600 times average sequence length). The run time of 1500 generations was chosen based on preliminary experiments (not shown) that demonstrated that, even with evolution still ongoing, the MADB is reliably established in

the first few hundred generations and tends not to change extensively after that time. To provide smoothing of momentary fluctuations in instruction concentration, the frequency of each of the 29 instructions was measured during each of the last 10 generations and averaged over these. We tested four different distributions of substitution probabilities: an increasing distribution (Experiment A), a decreasing distribution (Experiment B), one that increases and then decreases (Experiment C), as well as a randomly generated probability distribution (Experiment D). Note that, as the order of instructions is arbitrary, these four different substitution patterns have no inherent meaning. We performed 25 replicates of each experiment. All other parameters used in these experiments are the standard defaults described by Ofria and Wilke (2004), except that the standard fitness landscape of nine logic tasks (all distinct one- and two-input tasks) was replaced with the extended 73 logic tasks landscape, where all distinct logic tasks with up to three inputs are rewarded.

2.2. Mutation "spectra" experiments (Figure 4)

We created 18 different "spectra" that represent the relative frequency with which each instruction appears through mutation. For this experiment, we used the three manually constructed systematic variations of experiments A–C, while another 15 experiments used randomly generated mutation
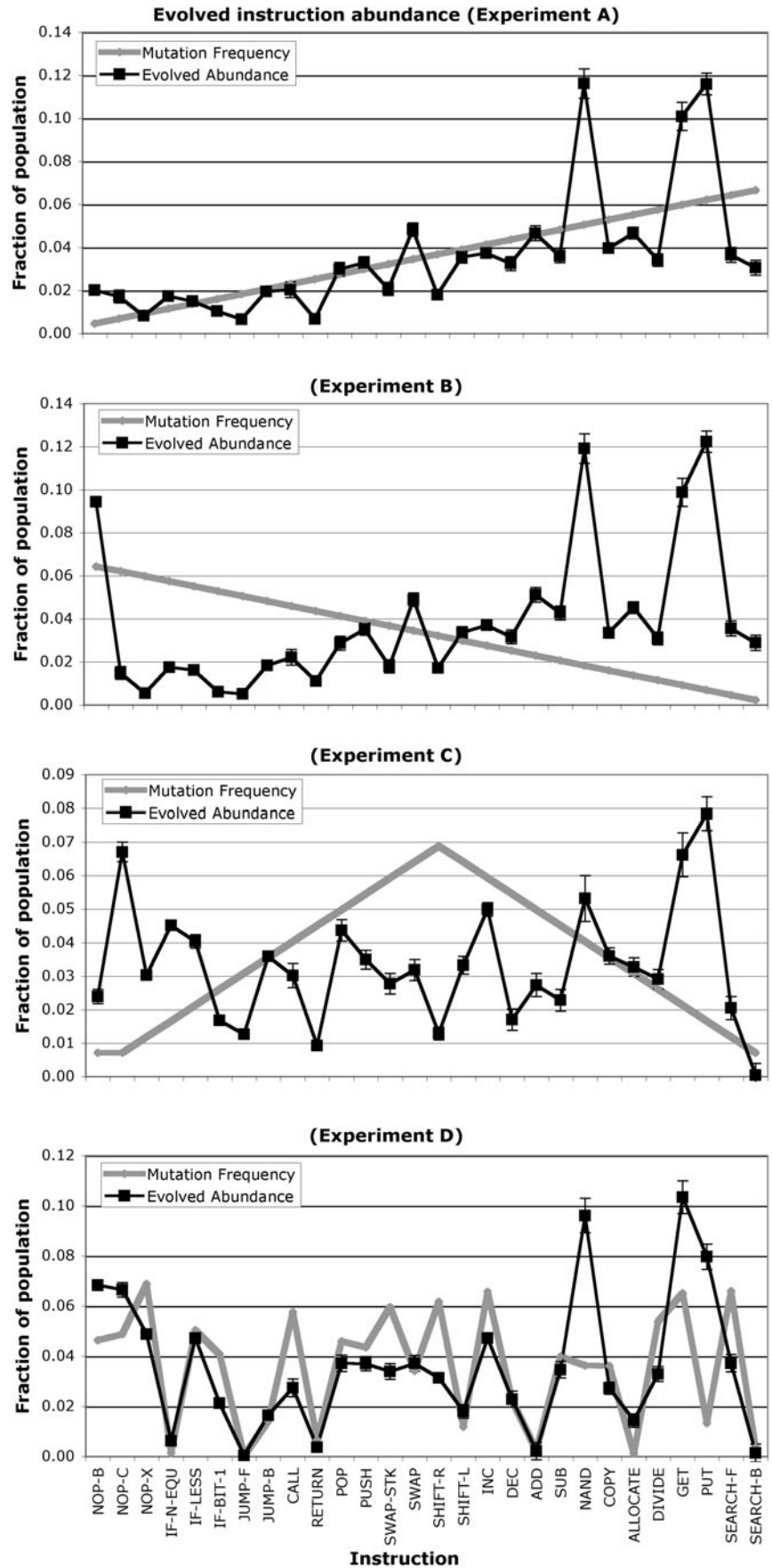
FIG. 3. Four experiments showing the evolved distribution of instructions as the underlying physics are changed. The gray curves represent the frequencies with which each instruction was presented to organisms through mutation, while black curves represent the relative abundances of the instructions in evolved populations. Each black curve represents the average of 25 populations (error bars are standard error). The general features of the selection-driven distribution are conserved even though the mutation frequencies are varied over large ranges. In experiments A, B, and C, the most frequent instructions appear 30 times more often than the least frequent. Experiment D shows one of 15 randomly generated distributions tested.
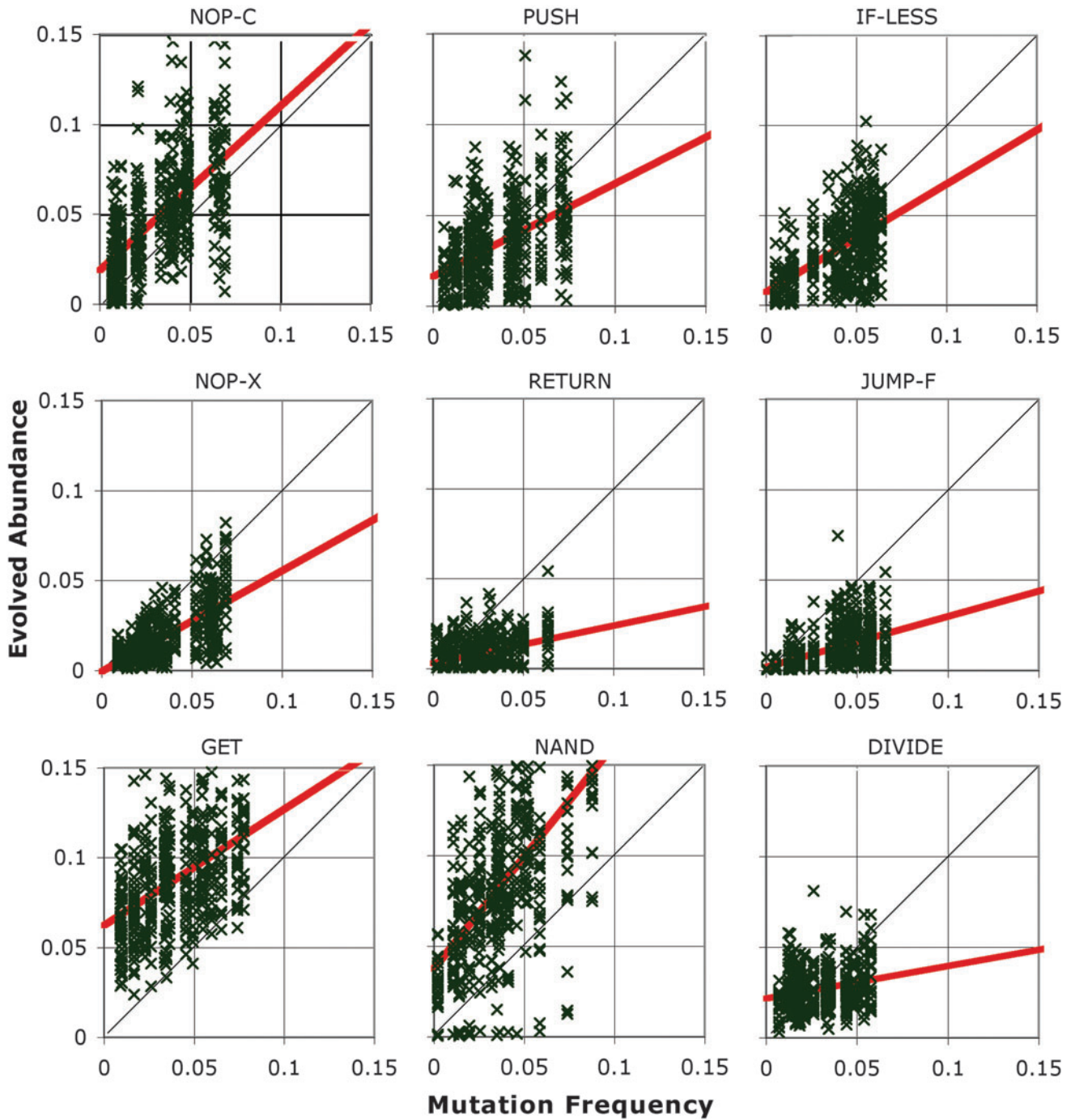
FIG. 4. The relationship between frequency of appearance in mutation and abundance in the terminal population for nine instructions. Trends above or below the unity line (gray) indicate a tendency to be selected for or against, respectively. GET typifies an instruction that frequently conveys a fitness advantage; it tends to be incorporated into genomes at a high level regardless of how often it appears as a mutation. Conversely, RETURN is not necessary for basic functions and is often lethal as a mutation, so it remains at a low level in nearly all runs, even when it appears frequently as a mutation. $N = 450$, 25 runs each using 18 different mutation frequency profiles. Trend lines (red) are a least-squares linear fit. The gray line is the unbiased assumption where the mutation frequency equals the evolved abundance. Axes are chosen such that this unbiased trend line is always a diagonal.

spectra, where each instruction was assigned a substitution probability bound between 0 and 0.08. This implies that any instruction that is assigned a vanishing substitution probability can never appear in the population. Each experiment was performed in 25 replicates.

2.3. Analysis

In our data analysis, we consider only 28 of the 29 instructions used in these Avida populations. One instruction, NOP-A, is used by the system as a temporary placeholder

to initialize empty memory in dividing organisms but is replaced as the organism copies its actual genotype into the empty array. Therefore, the number of NOP-A instructions appearing in the population fluctuates rapidly in a way that has no biological analogue and is highly dependent on the precise timing of the sample measurement. Also, this aspect of NOP-A cannot be selected for or against since the organisms do not have a choice about how empty memory is initialized. We therefore choose to exclude it from the analysis.

3. Results

Evolved Avida populations impress a distinctive pattern of instruction abundances onto their environment, and this pattern is largely conserved even when the availability of particular instructions is altered significantly. Figure 3 shows the average evolved abundances of the 28 instructions for the four different mutational profiles described in Methods. In each, the gray "mutational frequency" line represents the relative rates at which each instruction appears in mutation, and the black "evolved frequency" line represents the relative abundance of each instruction in the terminal population.

Certain features, such as the prominence of GET, PUT, and NAND, are conserved across the runs regardless of how their input (mutational) frequency is altered. This reflects the fitness benefit conveyed by these instructions, which are necessary for completing logic computations. In Avida, organisms are rewarded with increased processor time for successfully completing a variety of computational tasks, which play the role of exothermic catalytic reactions in the metabolism of digital organisms (Adami, 2006). In the landscape we used, 73 tasks are rewarded, each of which requires one to five NAND operations. Some other instructions, such as JUMP-F and RETURN, are frequently lethal when a mutation causes them to appear in the genome. As a result, they are rarely incorporated into genomes and appear underrepresented in the final population regardless of their mutational frequency. In their effect, such instructions are not unlike DNA codons that cause early termination of transcription in terrestrial biochemistry. A comparison between the four experimental treatments shows that the adaptive component dominates the mutational bias. Yet, the influences of chance and history (Travisano *et al.*, 1995; Wagenaar and Adami, 2004) are also present. For example, the relative abundance of the modifier instructions NOP-B and NOP-C differ in the different treatments, even though they could in principle substitute for each other. Even though the roles of the two instructions are similar, each instruction is assigned a particular functional role early on and can only be released from this role in very unlikely mutational events. This historical contingency is at the origin of many run-to-run differences between instruction abundances. But as long as the abundances differ significantly from the abiotic baseline, they still contribute to the MADB.

Figure 4 shows the relationships between mutation frequency and evolved population frequency for several instructions (the average relationship for all instructions is shown in Table 1). Each plot shows 450 data points, obtained from the 25 replicates of each of the 18 experiments where the mutational frequency spectrum was determined as described in Methods. In each plot, if selection did not constrain the organisms' use of each instruction, we would expect the data to fall on or near the unity line, reflecting that the organisms incorporated the instructions at the same rate at which they appear in mutation (*i.e.,* the rate at which they are "formed"). Data points above the unity line represent populations that used an instruction at higher than the expected rate, meaning that those instructions were preferentially incorporated into the evolving genomes. Points below the unity line represent instructions that were selected against. The slope of the distribution represents the extent to which an instruction's appearance in the final population depends on its availability in mutation.

While the data are widely distributed, some significant trends are evident. NOP-C, PUSH, and IF-LESS are close to neutral in average adaptive utility; they show broad distributions near the unity line. All three of these instructions exist in parallel with other instructions that can, to some degree, replace their function. NOP-C is part of a complementary set of address labels, including NOP-A and NOP-B, and an organism can construct labels using only a pair of NOP instructions if one is not available. PUSH is a stack operation, and organisms can function largely using register (as opposed to stack) storage if necessary. The flow-control

TABLE. 1. SELECTION BIASES [EQ. $\sim(1)$], SLOPE OF EVOLVED INSTRUCTION FREQUENCY vs. MUTATION FREQUENCY (SEE FIG. 4), VALUE OF THE *F*-TEST VARIABLE, AND PROBABILITY THAT THIS *F*-VALUE COULD HAVE BEEN THE RESULT OF CHANCE (*P*-VALUE), FOR THE 28 INSTRUCTIONS WE ANALYZED

Instruction	Selection bias	Slope ± standard error	F	P-value
PUT	31.1	0.83 ± 0.05	1,801	0
GET	21.6	0.65 ± 0.06	847	0
NAND	20.7	1.24 ± 0.07	416	0
NOP-B	12.6	0.94 ± 0.07	229	0
NOP-C	7.2	0.91 ± 0.06	93	0
INC	1.9	0.47 ± 0.04	80	0
SWAP	1.4	0.89 ± 0.05	10.2	4.56×10^{-5}
ALLOCATE	1.2	0.55 ± 0.03	159	0
ADD	0.2	0.75 ± 0.05	12.4	5.74×10^{-6}
PUSH	-0.4	0.51 ± 0.05	50	0
POP	-1.1	0.65 ± 0.05	33	3.8×10^{-14}
DIVIDE	-1.2	0.18 ± 0.04	285	0
SHIFT-L	-1.6	0.55 ± 0.03	139	0
SEARCH-F	-2.6	0.36 ± 0.04	173	0
IF-N-EQU	-3.0	0.57 ± 0.03	144	0
JUMP-B	-3.4	0.25 ± 0.02	731	0
COPY	-3.6	0.30 ± 0.04	231	0
IF-LESS	-3.9	0.60 ± 0.05	95	0
DEC	-4.3	0.62 ± 0.04	128	0
SUB	-4.8	0.71 ± 0.04	106	0
CALL	-5.2	0.39 ± 0.06	132	0
SWAP-STK	-7.0	0.45 ± 0.03	450	0
RETURN	-7.4	0.21 ± 0.02	1562	0
NOP-X	-7.6	0.56 ± 0.03	539	0
IF-BIT-1	-9.1	0.36 ± 0.03	940	0
SEARCH-B	-10.2	0.33 ± 0.02	1,477	0
SHIFT-R	-10.4	0.48 ± 0.02	1,137	0
JUMP-F	-11.1	0.28 ± 0.02	1,751	0

instruction IF-LESS can be entirely replaced by IF-N-EQU, and to some degree with IF-BIT-1, if it is not available.

NOP-X, RETURN, and JUMP-F are examples of maladaptive instructions that are selected against, with most of their instances appearing below the unity line. NOP-X is a neutral but nonfunctional operation that merely consumes a single CPU cycle. While it does not harm an organism, if inserted into a loop it can cause a significant delay in the time required for a genome to complete tasks and reproduce. We see it somewhat selected against. RETURN and JUMP-F, however, are flow control instructions that are both unnecessary (their functions can be completely replaced by other flow instructions like JUMP-B) and generally maladaptive. When a JUMP-F or RETURN is inserted or substituted into a genome, the resulting genome will often skip large blocks of instructions or enter infinite loops; such mutations are usually fatal. As a result, these are the two most strongly selected-against instructions.

Obviously, beneficial instructions include GET and NAND, which are essential for the completion of mathematical tasks. GET and PUT (not shown in Fig. 4) are responsible for input and output within the computational metabolism; genomes cannot gain any advantage over the ancestor without using them. As a consequence they are strongly selected for. NAND is also present in high abundance in nearly all populations, as it is used to perform computations on the input stream accessed via GET (NAND is the only instruction available to compute logic tasks and, therefore, must be used increasingly in order to achieve higher fitness). However, it shows a strong—in fact greater than unity—dependence on the input frequency of mutation. When NAND is produced more often, tasks evolve more quickly, and as a consequence NAND instructions accumulate in the sequence, which leads to an evolved abundance that exceeds the production frequency.

DIVIDE is an interesting case. It has the least dependence on mutation rate of any instruction, but even at near-zero mutation rate it appears at a significant fraction of the population, shown by the high intercept of the trend line relative to maladaptive instructions like NOP-X and JUMP-F. DIVIDE splits a genome in half and is essential for the reproduction of organisms after they have copied all their instructions. However, if it appears in an inappropriate location, the organism will divide prematurely; this is nearly always fatal. As such, DIVIDE almost always appears exactly once per genome regardless of mutation effects. The vertical scatter seen in the DIVIDE frequency is largely due to variation in the length of the evolved genomes.

Table 1 gives full results for the selection bias of the 28 instructions we analyzed. Selection bias for each instruction is an approximation of the tendency of selection pressures to elevate (or suppress) the population concentration of a single instruction. The selection bias SB_i for instruction i is computed via

$$SB_i = \sum_n C_{i,n} - \mu_{i,n} \tag{1}$$

where $C_{i,n}$ is the fractional population concentration of instruction i in trial n and $\mu_{i,n}$ is the probability of instruction i appearing via mutation in trial n. This bias can be positive or negative. The instructions in Table 1 are sorted by this measure (instructions with the strongest selection bias are at

the top). If an instruction is incorporated into the genome at the same frequency at which it appears in mutation, we expect SB to be zero. We also list the slope of the linear fit for each instruction, which indicates the degree to which an instruction's population abundance depends on its mutation frequency. To test whether the linear fit can significantly distinguish the abundance distribution of the instruction from the unbiased assumption (the line with unit slope and vanishing intercept), we conducted a test of variances (F-test) and listed the value in Table 1, along with the associated P-value. According to this test, all instructions deviate significantly from the neutral evolution assumption.

4. Discussion

Artificial life is a useful tool for astrobiology, in that it can examine the fundamental processes of life with an eye toward identifying universal phenomena: features of life that may be detectable regardless of a life-form's substrate or particular form. It may be seen, therefore, as an approach toward solving the "single data point" problem, that is, that we know only one example of evolved life (the terrestrial biosphere) and, therefore, cannot draw scientific conclusions about the universality of features we observe. Using artificial-life techniques, we can test conjectures about observable invariants of life; other examples include measuring the reduction of local entropy induced by cellular automata in an artificial chemistry (Centler et al., 2003).

We have demonstrated the repeatability of the MADB in populations of avidians and characterized the signature's robustness with respect to alterations in the underlying physics. We found that, although significant variations in monomer abundance patterns do appear as evolutionary experiments are repeated, general features (such as the selection for mathematics instructions and the suppression of frequently lethal flow-control instructions) are conserved. This feature mirrors observations of functional and structural constraints on the composition of proteins in the terrestrial biosphere, where for example hydrophobicity or stability requirements constrain the type, but not the identity, of amino acids incorporated into proteins (Forsdyke, 2005).

More importantly, in no case does the evolved abundance pattern ever resemble the pattern predicted by the system's physics, which is the most important characteristic of a life diagnostic or biosignature. This robustness clearly derives from evolutionary necessity. The organisms' metabolism and composition are subject to selection for fitness, and the features of that composition will be impressed upon their environment. This observation also has a parallel in terrestrial biochemistry, where mutational constraints (giving rise to GC bias) can influence the amino acid composition of a protein (Gu et al., 1998; Singer and Hickey, 2000) but cannot change the basic pattern of relative residue abundances into those observed in abiotic samples.

This conclusion indicates one possible direction for the search for non-terrestrial life via a method that is agnostic of terrestrial biochemistry, that is, "non-Earth-centric" life detection. By modeling or recording the range of plausible abiotic formation ratios of various chemical compounds, we may examine samples for compounds appearing outside those ranges. Measurements of chemical concentrations that

deviate from those ranges may indicate that an evolved metabolism is selectively synthesizing useful compounds. On the contrary, because diagenesis can obscure a biological signature if enough time has passed since its deposition, the absence of a detectable signature is not necessarily conclusive for the absence of historical life.

It is important to recognize the distinction between the formation of the MADB via biosynthesis and its detectability in an environmental sample. If the biotic signature was laid down long before observation, diagenesis could obscure the MADB because it is possible that different monomers degrade at different rates. For example, when natural sediments degrade over time, amino acids of low molecular weight can become predominant simply because they are more stable (Abelson and Hare, 1969; Elster *et al.,* 1991). In addition, the signature will only be measurable if the population of organisms is sufficient to generate measurable biomass with respect to the background chemistry. Since we examined only the composition of the avidians themselves, this study is modeling only the formation and evolution of the signature, not whether it would be measurable in the environment. We refer the reader to our prior work (Dorn *et al.,* 2011) for further discussion of this issue.

Understandably, there are considerable practical limitations and obstacles to determining the baseline distribution to which we would compare an observed pattern of molecules. For example, the ability of this strategy to reject false positives depends critically on our ability to thoroughly characterize the range of possible abiotic distributions in advance. Such an endeavor should begin with the most exhaustive experimental analysis possible of monomer formation in all conceivable abiotic conditions, and we should remain vigilant to the possibility that unconsidered special cases (*i.e.,* unusual combinations of local environmental chemistry, temperature, radiation, or other factors) might include dynamics that produce an unexpected abundance pattern. While theoretical modeling and numerical simulation of planetary atmospheres and surface chemistry are important components of establishing the baseline, if possible they should be supplemented by simulations in a laboratory. For example, the geochemistry of exoplanets may be modeled by using specific geochemical cycles to constrain spectral signatures (Kaltenegger and Sasselov, 2010), while Mars' evaporite geochemistry is readily simulated under martian environmental conditions in the laboratory (Moore *et al.,* 2010). In some instances, the knowledge of the background distribution does not have to be precise, such as when abiotic chemistry predicts a smooth distribution of polymer abundances, while the biotic distribution is discrete [*cf.* the Lego Principle of Davies *et al.* (2009) and McKay (2004)]. Once a candidate set of polymers has been identified and a baseline distribution suggested, planetary missions could be designed that measure thousands of chemical relative abundances by using, for example, on-chip liquid chromatography or other methods such as Raman spectroscopy or UV fluorescence (McKay, 2010). However, the technology to perform high-throughput targeted relative abundance measurements is probably still years away. Given such measurements, however, standard machine-learning techniques can be applied to distinguish biotic from abiotic patterns (Dorn *et al.,* 2003; Dorn, 2005).

Yet, even in the light of such practical difficulties, we see the MADB as an important means to simulate and predict life's signature because it is the inevitable result of a fundamental life process (evolutionary selection) that is completely independent of information about any specific biochemistry. If the abiotic distribution is characterizable, the MADB should be detectable as long as the life-form under study employs in its metabolism any members of the chemical family under study.

Acknowledgments

Author Disclosure Statement

The authors state that no competing financial interests exist.

Abbreviation

MADB, Monomer Abundance Distribution Biosignature.

References

Abelson, P.H. and Hare, P.E. (1969) Recent amino acids in the Gunflint Chert. *Year B Carnegie Inst Wash* 69:208–210.

Adami, C. (2006) Digital genetics: unravelling the genetic basis of evolution. *Nat Rev Genet* 7:109–118.

Bains, W. (2004) Many chemistries could be used to build living systems. *Astrobiology* 4:137–167.

Benner, S.A. and Switzer, C.Y. (1999) Chance and necessity in biomolecular chemistry: is life as we know it universal? *Simplicity and Complexity in Proteins and Nucleic Acids,* edited by H. Frauenfelder, J. Deisenhofer, and P.G. Wolynes, Dahlem University Press, Berlin, pp 339–363.

Benner, S.E. (2010) Chemistry, life, and the search for aliens. *Proc SPIE* 7819, doi:10.1117/12.866981.

Centler, F., Dittrich, P., Ku, L., Matsumaru, N., Pfaffmann, J., and Zauner, K.P. (2003) Artificial life as an aid to astrobiology: testing life-seeking techniques. In *Advances in Artificial Life— Proceedings of the 7th European Conference on Artificial Life (ECAL),* Lecture Notes in Artificial Intelligence Vol. 2801, edited by W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, and J. Ziegler, Springer, Berlin, pp 31–40.

Davies, P.C.W., Benner, S.A., Cleland, C.E., Lineweaver, C.H., McKay, C.P., and Wolfe-Simon, F. (2009) Signatures of a shadow biosphere. *Astrobiology* 9:241–249.

Dorn, E.D. (2005) Universal biosignatures for the detection of life. Ph.D. Thesis, California Institute of Technology, Pasadena, CA.

Dorn, E.D., McDonald, G.D., Storrie-Lombardi, M.C., and Nealson, K.H. (2003) Principal component analysis and neural networks for detection of amino acid biosignatures. *Icarus* 166:403–409.

Dorn, E.D., Nealson, K., and Adami, C. (2011) Monomer abundance distribution patterns as a universal biosignature: examples from terrestrial and digital life. *J Mol Evol* 72: 283–295.

Elster, H., Emanuel, G., and Weiner, S. (1991) Amino acid racemization of fossil bone. *J Archaeol Sci* 18:605–617.

Forsdyke, D.R. (2005) Functional constraint and molecular evolution. In *Encyclopedia of Life Sciences,* John Wiley & Sons, Chichester, UK, www.els.net.

Gu, X., Hewett-Emmett, D., and Li, W.H. (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102–103:383–391.

Joshi, A., Castillo, R.B., and Mueller, L.D. (2003) The contribution of ancestry, chance, and past and ongoing selection to adaptive evolution. *J Genet* 82:147–162.

Kaltenegger, L. and Sasselov, D. (2010) Detecting planetary geochemical cycles on exoplanets: atmospheric signatures and the case of SO_2. *Astrophys J* 708:1162–1167.

Kvenvolden, K.A., Lawless, J.G., and Ponnamperuma, C. (1971) Nonprotein amino acids in Murchison Meteorite. *Proc Natl Acad Sci USA* 68:486–490.

Lovelock, J.E. (1965) A physical basis for life detection experiments. *Nature* 207:568–570.

McDonald, G.D., Thompson, W.R., Heinrich, M., Khare, B.N., and Sagan, C. (1994) Chemical investigation of Titan and Triton tholins. *Icarus* 108:137–145.

McKay, C.P. (2002) Planetary protection for a Europa surface sample return: the Ice Clipper mission. *Adv Space Res* 30:1601–1605.

McKay, C.P. (2004) What is life—and how do we search for it in other worlds? *PLoS Biol* 2:1260–1263.

McKay, C.P. (2010) An origin of life on Mars. *Cold Spring Harb Perspect Biol* 2:a003509.

Miller, S.L. (1953) A production of amino acids under possible primitive Earth conditions. *Science* 117:528–529.

Miller, S.L. (1957) The formation of organic compounds on the primitive Earth. *Ann NY Acad Sci* 69:260–275.

Moore, J.M., Bullock, M.A., Newsom, H., and Nelson, M. (2010) Laboratory simulations of Mars evaporite geochemistry. *J Geophys Res* 115, doi:10.1029/2008JE003208.

Munoz-Caro, G.M., Meierhenrich, U.J., Schutte, W.A., Barbier, B., Segovia, A.A., Rosenbauer, H., Thiemann, W.H.P., Brack, A., and Greenberg, J.M. (2002) Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature* 416:403–406.

Ofria, C. and Wilke, C.O. (2004) Avida: a software platform for research in computational evolutionary biology. *Artif Life* 10:191–229.

Shapiro, R. and Schulze-Makuch, D. (2009) The search for alien life in our solar system: strategies and priorities. *Astrobiology* 9:335–343.

Singer, G.A. and Hickey, D.A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* 17:1581–1588.

Summons, R., Albrech, P., McDonald, G., and Moldowan, J. (2008) Molecular biosignatures. *Space Sci Rev* 135:133–159.

Travisano, M., Mongold, J.A., Bennett, A.F., and Lenski, R.E. (1995) Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* 267:87–90.

Wagenaar, D.A. and Adami, C. (2004) Influence of chance, history, and adaptation on digital evolution. *Artif Life* 10:181–190.

Ward, P.D. and Benner, S.A. (2007) Alien biochemistries. In *Planets and Life: The Emerging Science of Astrobiology,* edited by W.T. Sullivan III and J.A. Baross, Cambridge University Press, Cambridge, pp 537–544.

Wolman, Y., Haverlan, W.J., and Miller, S.L. (1972) Non-protein amino acids from spark discharges and their comparison with Murchison Meteorite amino acids. *Proc Natl Acad Sci USA* 69:809–811.

Address correspondence to:
Christoph Adami
Microbiology and Molecular Genetics
Michigan State University
East Lansing, MI 48824

E-mail: adami@msu.edu