Universal Biosignatures for the Detection of Life

Thesis by

Evan D. Dorn

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2005
(Submitted May 27, 2005)

To my parents, Mrs. Theresa C. M. Dorn and Dr. Ronald V. Dorn III, whose support and encouragement has gotten me here.

Acknowledgements

I am indebted most to my advisors, Christoph Adami and Kenneth H. Nealson, for their support and guidance during my studies. Thanks as well to the members of my committee who have evaluated this thesis: Dmitri Psaltis, Joseph Kirshvink, and Christof Koch.

Many researchers have contributed to this work. Gene McDonald and Michael C. Storrie-Lombardi did the early work in amino-acid patterns that started me on this path. Co-workers, collaborators, and friends at CIT, JPL, and elsewhere have contributed discussions and analysis that helped shape my research. Among them are Pamela Conrad, Sasha Tsapin, Misha Goldfeld, Claus Wilke, Robert Forster, Charles Ofria, D. Allan Drummond, Alan Hampton, Jesse Bloom, and Kaben Nanlohy.

Friends and family members have helped with support and with the titanic tasks of editing and proofreading over the years, and to them I extend my eternal gratitude: Ronald V. Dorn III, Jessy Dorn, Tinsley Davis, Scott Van Essen, Diana Sherman, Joseph Cook, and Melinda Epler.

My heartfelt thanks and apologies to the myriad others who have contributed in small ways or whose names I have forgotten.

Some of the research described here was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA), with support from the Director's Research and Development Fund (DRDF) and from the National Science Foundation under contract No. DEB-9981397.

Abstract

My goal is to identify processes of life that leave measurable effects on an organism's environment, but which are not tied to any particular biochemistry, in order to build a conceptual framework for the search for extraterrestrial life. To this end, I test a pair of phenomena that appear in both terrestrial (biochemical) life and in digital life. Because these two life forms are different and unrelated, any phenomenon measurable in both is suggested to be universal.

The Monomer Abundance Distribution Biosignature (MADB) is any measurement of the relative concentrations of related chemical compounds that cannot be explained by abiotic processes. I observe that living systems synthesize specific chemical compounds at rates that maximize their fitness. As a result, life-bearing environmental samples exhibit compounds in abundance ratios that are clearly not the result of abiotic synthesis because those ratios belie the formation kinetics and thermodynamics that would constrain abiotic synthesis. Often, biotic samples contain high concentrations of specific large, complex molecules that are never seen in abiotic synthesis and cannot be explained unless highly specific catalysts (i.e., enzymes) are present, and energy is expended to drive thermodynamically unfavorable reactions. I catalog this effect as it appears in terrestrial biochemical systems, including amino acids and carboxylic acids, and demonstrate the universality of selection's action on the monomeric composition of life forms by studying analogous examples in digital life. I suggest how this phenomenon provides a route to the detection of even unusual or unforeseen biochemistries, and give examples of detection methods using pattern-recognition techniques that may allow us to empower an autonomous system with the general ability to detect life forms.

The Layered Trophic Residue Biosignature (LTRB) is any observation of stratification in solute chemistry that indicates metabolic activity by a sequence of diverse communities. When multiple chemical resources are available, natural selection drives adaptive radiation and the formation of specialist phenotypes. Competition ensures that specialists consume resources in decreasing order of energetic potential when resources diffuse through a medium near a boundary. The result is strata of chemicals appearing in order of redox potential, which is best explained by the presence of life.

Contents

III Detection of the Monomer Abundance Distribution

List of Figures

List of Tables

Part I

Introduction and Summary

Chapter 1

Introduction

1.1 The search for life

The potential existence of life outside the terrestrial biosphere remains one of the most important unanswered questions in science. To date, the NASA Viking landers of 1971 and 1972 have conducted the only in-situ experiments dedicated to detecting life on another planet. While most scientists believe the tests were negative, the interpretation of the results remains in dispute today.

Many astrobiologists feel that the search for extraterrestrial life should be "non-Earth-centric," meaning that our experimental methods should be minimally biased by our present understanding of terrestrial forms of life. We should not, for example, depend on an experiment which searches extraterrestrial soil for DNA, in case a putative organism uses a different molecule for information storage. More extreme suggestions imply that we should not even assume that life forms are based on organic chemistry (some call such an assumption "carbon chauvinism"). While we can debate the likelihood of such alternative biochemistries, our search should allow for the maximum number of possibilities.

Leaving our potential space of biochemistries deliberately broad, however, begs the question of what exactly we are looking for; is it possible to construct an experiment to look for for an unknown biochemistry?

I argue in this thesis that the fundamental nature of living things guarantees they will alter their environment in ways that should be detectable through geochemical analysis. Furthermore, I suggest two general classes of biosignatures that may serve our purpose, and propose a strategy for detecting one of them in a non-Earth-centric way.

If the goal is the detection of life without foreknowledge of, or bias about, its chemical makeup, then we must start by identifying features of life that we believe are common to living systems. Without defending any particular definition of life (a contentious and risky proposition in any scientific audience), I suggest that most biologists would allow that living systems metabolize, i.e., couple energy-producing reactions to energy-consuming reactions in order to drive biosynthesis, and that living systems evolve and are subject to selection. Indeed, many biologists consider metabolism to be an important part of the definition of life.

Metabolism, constrained by natural selection, leads to two phenomena which have been measured in terrestrial geochemistry. Analysis of the first, which I call the monomer abundance distribution biosignature (MADB), constitutes the bulk of the work presented in this thesis. Preliminary results concerning the second, dubbed the layered trophic residue biosignature, are also presented.

1.2 The monomer abundance distribution biosignature

The space of possible chemical environments in the absence of life is large, but finite, and we may be able to make exclusionary statements about it. Given comparable molecules such as similar ligands from a common chemical family, we may be able to make confident statements about their relative concentration. For example, in the atmosphere of a planet whose crust contains abundant iron or other metals, we should not see a high quantity of molecular oxygen relative to less-reactive species like carbon dioxide and nitrogen because oxygen reacts quickly with metals to form mineral oxides. Consider monosaccharides (glucose, fructose, etc.) as a second example: all have the same formula, similar structure, and similar formation energy. The observation of an extremely high concentration of any one monosaccharide species relative to the others would indicate that some other phenomenon was actively selecting for the increased concentration of that compound. The consideration of *relative* (as opposed to absolute) concentration is critical: it insulates our analysis from problems like dilution, dispersion, and/or the general availability of carbon in the environment.

I use amino acids as an example in all of my research, because a wealth of data exist. In an environment conducive to the formation of organic molecules and protected from destructive influences like UV light or strong oxidants, the concentration of heavy amino acids like arginine [($C_6H_{14}N_4O_2$), molecular weight 174.20)] and phenylalanine [($C_9H_{11}NO_2$), molecular weight 165.19)] should be low relative to simpler compounds like glycine [($C_2H_5NO_2$), molecular weight 75.07)]. The large number of synthesis steps, high kinetic barriers, and many conformers of the heavier molecules make it extremely unlikely that they will be formed in large quantities unless catalysts are available which increase the formation rates of those specific compounds. Indeed, fifty years of experimentation in abiotic formation of amino acids from many different precursors and in a variety of environments and energy sources have repeatedly generated mixtures of amino acids dominated primarily by glycine with a few other low-weight constituents such as alanine, beta-alanine, and sarcosine. [51, 59, 60, 62, 64, 89, 95].

The manufacture of catalysts to lower kinetic barriers and increase the rate of specific reactions, and indeed, the coupling of energy-producing reactions to active catalysts in order to drive otherwise thermodynamically unfavorable reactions are two of the primary functions of metabolism in living systems. Terrestrial biosystems use protein enzymes for this job. While a hypothetical extraterrestrial (ET) life form might not employ proteins as we understand them, we may safely presume that it would nonetheless require some form of cellular machinery to conduct its metabolic chemistry.

Given that any hypothetical biochemistry would include cellular machinery to amplify the rates of necessary reactions, I therefore contend that selection will act to make that machinery highly *specific*. Organisms which synthesize their component compounds at ideal rates will be more fit than those which waste energy by synthesizing too much of any one compound, or which unnecessarily constrain their reproduction by synthesizing too little of another key ingredient.

We expect the component molecules of organisms and their waste products to be present in the environment, subject of course to diagenesis and breakdown. These compounds, then, will stand out from the "background" chemistry we would expect if life were not present. In practice, we expect this effect to be easiest to observe when we make simultaneous measurements of multiple compounds from related families

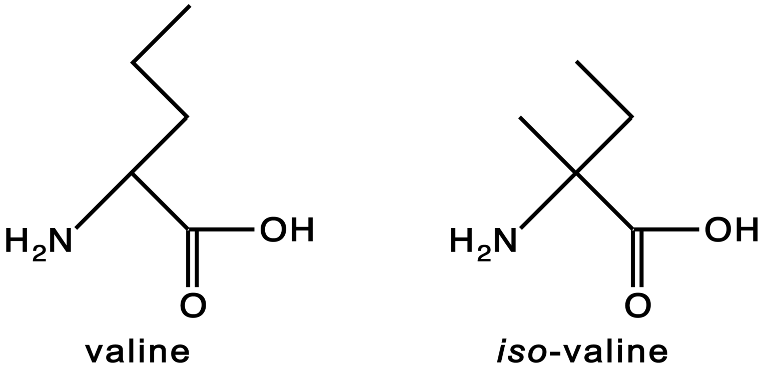


Figure 1.1: The molecular structures of valine and isovaline

Sample	**Source**	**Ratio**
Murchison	Engel and Nagy, 1982	0.36
Murchison	Engel and Nagy, 1982	5.45
ALHA 77306	Cronin et al. 1979	0.67
ALHA 77306	Cronin et al. 1979	0.50

Table 1.1: Valine:isovaline ratios from meteorite samples

and compare their abundance ratios.

For example, consider the conformers valine and isovaline, which have identical molecular weights and similar structure differing only in the position of one methyl group. (Figure 1.1) Because of these similarities, we would expect their appearance as a result of abiotic synthesis to be similar. This in fact is the case: in the few studies of uncontaminated meteorite amino acid contents which have measured and reported both valine and isovaline [37,27], their measured concentrations have always been within an order of magnitude as shown in Table 1.2. In the biosphere, however, where valine is synthesized rapidly along with the other protein amino acids in very large quantities that dwarf isovaline, the ratio is effectively infinite. Indeed, the protein amino acids so thoroughly dominate terrestrial amino acid profiles that few non-protein amino acids are ever reported at all in the numerous studies I have considered; beta-alanine is the only exception I have observed in a search of available literature.

This dominance by the protein amino acids is evidence of the selective pressure I have previously mentioned: once the amino acid language of protein formation was established in early evolution, selection favored those organisms that efficiently catalized the synthesis of the protein amino acids most efficiently.

I hypothesize, then, that we may observe the presence of life in abundance ratios of related compounds that cannot be explained by the kinetics and thermodynamics of abiotic synthesis, and I call this effect the monomer abundance distribution biosignature (MAD Biosignature or MADB). Put simply, the rationale behind the MADB may be described this way: abiotic synthesis forms the compounds that are *easy to make*. Organisms, on the other hand, make the compounds *that they need* in order to survive and compete. These two fundamentally different constraints should produce different abundance ratios.

We would like to consider more than just a pair of molecules. If we measure the concentrations of a set of several related chemicals, the overall pattern of that group constitutes a *profile* which may be diagnostic for the underlying formation mechanics of that family of compounds. Consider the frequency of letters in written English as a visualization tool by analogy. If the letters did not carry meaning or were selected by an "abiotic" process, we would expect all twenty-six letters to appear with the same frequency. However, letters have function in written language: they form words, which carry meaning. Their various utility in this function dictates that they appear in specific frequencies. These frequencies are known. Figure 1.2 gives two examples of the distribution spectra of letters in English texts: chapter 1 of *The Origin of Species* by Charles Darwin, and the book of Genesis from the King James Bible. The letter frequencies reflect, for example, that vowel sounds are necessary in every syllable, but only five of the letters are primary vowels; as a result, there are four primary vowels among the five most common letters. This pattern is used for a task conceptually similar to our task of detecting life: letter frequencies are one of the primary tools of classical cryptanalysis, the science of elucidating meaning in an apparently meaningless or random ciphertext.

In much the same way that certain distributions of letter frequency can indicate the presence of meaning in a sequence of characters, patterns of concentration in chemical families in terrestrial soil samples can indicate the presence of the action of

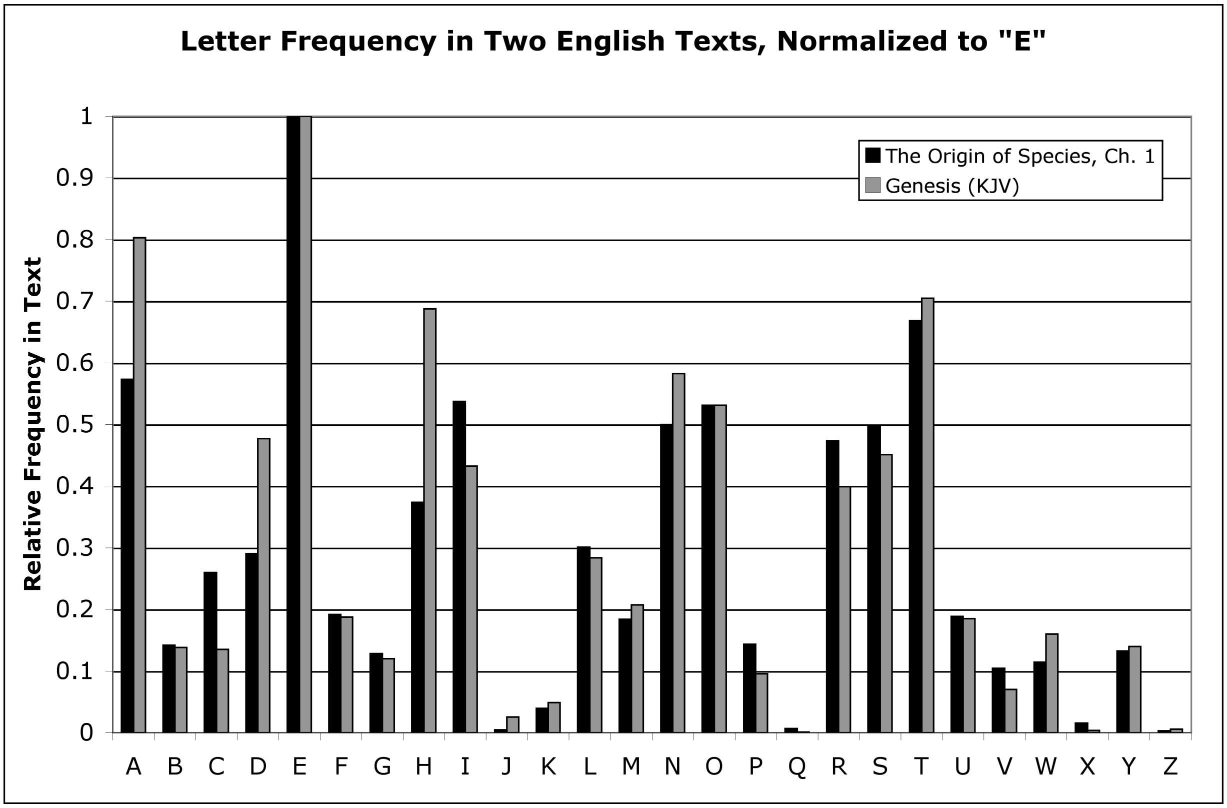


Figure 1.2: The relative distribution of the letters of the Roman alphabet in two representative samples of English text.

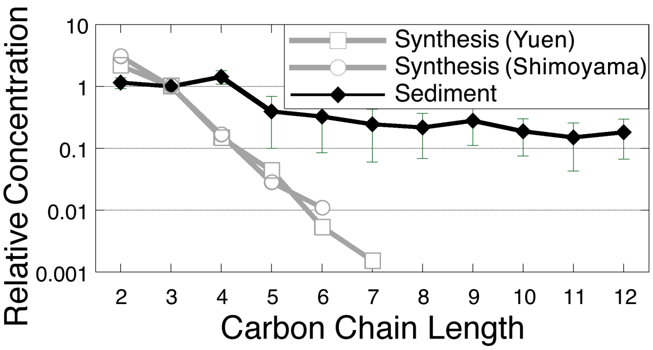


Figure 1.3: Short-chain carboxylic acids from terrestrial sediments compared to those from two spark-synthesis experiments. Sediment data is the average of twelve measurements from [81]. Error bars are one standard deviation.

selection, and therefore, life. Figure 1.2 compares the the relative concentrations of twelve low-weight, straight-chain carboxylic acids (C2:0, acetic acid, through C12:0, dodecanoic acid), as measured in terrestrial sediments [81] and from extracts of two spark-synthesis experiments [80, 97]. Note how the concentrations from extracts of abiotic syntheses drop off rapidly with increasing molecular weight: the larger molecules have higher formation energies, more conformers, and require a greater number of synthesis steps, so with each additional methyl group, they become exponentially less likely to form. The acids extracted from terrestrial sediments, which contain life, do not conform to the abiotic form. Instead, the heavier acids are present in abundance, presumably because they serve a biological function.

More rigorously, if we consider the concentration c of a particular chemical in an environmental sample, we expect that concentration to be the result of some complex function involving the rates of formation k_f and breakdown or removal k_b from both biotic and abiotic sources:

$$c = F(k_{f,abiotic}, k_{f,biotic}, k_{b,abiotic}, k_{b,biotic}, \dots) \tag{1.1}$$

When the biologically mediated terms $k_{b,abiotic}$ and $k_{b,biotic}$ are nonzero, it is possible that c may take on a different range of values than is possible in the purely abiotic case. In general, this function is not knowable, and other inputs such as overall dilution can conspire to make the particular value of an individual c useless in the general case. Instead, we consider M simultaneous measurements of several chemical species $c_1, \ldots, c_M$ as the vector $\boldsymbol{c}$, which we normalize either by computing a unit vector or by dividing the vector by a particular component c_M. Then, a change in the abundance ratio of any two chemicals $c_i : c_j$, $c_i : c_k$, etc., is equivalent to a change in the direction of the vector $\boldsymbol{c}$. We can then redefine the MADB in these terms: given a measurement vector $\boldsymbol{c}$ of several related compounds in an environmental sample, our sample contains life if the value of $\boldsymbol{c}$ represents the presence of these compounds in proportions that cannot be explained by abiotic processes.

The MAD Biosignature, then, meets the criteria for non-Earth-centric life detection: by looking for patterns which fall outside the range of possible abiotic equilibria, we can discover life regardless of its biochemical similarity to our own. This means, of course, that for such a survey to be meaningful and to have a good chance of detecting life, we must satisfy two conditions. First, the system making the analysis must analyze a sufficient source of data about environmental chemistry (i.e., a sufficient number of compound families and a sufficient number of examples of each family) that it will actually measure at least some of the compounds used by the putative life form it wishes to discover. Second, we must carefully and exhaustively characterize the range of possible abundance patterns which we consider abiotic.

The latter statement leads to an important conclusion about the search for extraterrestrial life: if we are to remain open-minded about the composition of possible biota, we best prepare for the search not by studying life, but by studying the absence of life.

1.3 A brief comment on chirality

Homochirality, particularly of amino acids, is an often-cited potential biosignature [56]. I have chosen not to involve chirality in my analysis because I would like to make as few assumptions as possible about the nature of possible ET life forms. The analysis, therefore, is left open to the possibility that a biochemistry might exist in which both enantiomers of optically active compounds are used in equal quantity. Furthermore, the analysis performed herein demonstrates that in known systems, monomer abundance profiles made without consideration to optical activity are sufficient to distinguish biotic and abiotic systems.

However, the homochirality of terrestrial amino acid biosynthesis is a particularly lucid example of the primary argument behind the MADB: given achiral or racemic precursors, enantiomers have exactly equal formation kinetics and thermodynamics, and in any abiotic synthesis a racemic mixture should be formed. The drastic difference in concentration between enantiomers of amino acids in the terrestrial biosphere is impossible to explain by mere thermokinetics and is therefore clear evidence that a selection pressure — the necessity of consistent peptides for the manufacture of functional proteins — is affecting synthesis.

1.4 A comment on experimental design for missions

While I use quantified measurements of individual monomers in all sections of this thesis, that choice is in large part dictated by the data available in the literature. Others have correctly pointed out that actual quantification of individual amino acids, for example, would be a difficult and probably unnecessary task for an automated spacecraft.

The information about the distributions of monomers is also contained in the raw output of whatever instrument is used to make the measurement. In all the monomer profiles I have used here, the quantification step involves at least one form of gas- or liquid-phase chromatography and/or mass spectrometry to separate individual monomers after the desired chemical family has been separated from the bulk sample via other chemical methods. The raw chromatographic spectra contain at least all of the information that is in the tables of concentrations extracted from those spectra.

The pattern-recognition methods I use in Chapter 6 can be easily adapted to any sort of input data, including chromatographic spectra. In a mission, it would likely make more sense to train the pattern-recognition system on raw (abiotic) spectra from an instrument included in the spacecraft such as a gas chromatograph-mass spectrometer (GCMS) like the equipment included on the Viking and Huygens landers. The pattern-recognition system would then be looking for patterns of chromatographic peaks which deviated significantly from its database of “allowable” abiotic spectra. Using raw spectra as the abiotic training examples would both reduce the complexity of the chemical analysis required to be performed on an unknown sample before the data could be delivered to the pattern-recognition system, and would allow for the possibility of detection of entirely unanticipated chemical families. If, for example, the often-hypothesized “silicon-based biochemistry” were present, the anomalous silane peaks in a chromatographic spectrum could trigger a positive result, even if silanes were never considered during the research into possible abiotic environments.

Part II

Formation of the Monomer Abundance Distribution Biosignature

Part II comprises three chapters: two representing work that has been submitted to peer-reviewed journals for publication, and one representing unpublished results that extend those results.

Herein I define the monomer abundance distribution biosignature, and consider its appearance in two families of terrestrial biochemicals and one example of non-terrestrial life.

In Chapter 2, I draw first on a history of research into prebiotic organic synthesis conducted between 1953 and 2004. Studies of chemical synthesis have yielded amino acids and/or carboxylic acids from conditions intended to model the prebiotic terrestrial environment, planetary bodies including Jupiter, Titan, and Triton, and interstellar ice particles and gas.

I compare the data from these laboratory studies to extractions of monomers from another putative source of abiotic chemical synthesis: carbonaceous chondrites, particularly the Murchison meteorite, which are believed to have been formed during the early history of the solar system and that contain organic compounds believed to date from that period.

These abiotic sources are contrasted to samples of the same chemical families from the terrestrial biosphere. I draw on a representative sample of a vast literature of amino acid and carboxylic acid profile studies from the Earth. — This study goes on to demonstrate that the same phenomenon that causes biotic profiles to differ from abiotic profiles — selection – also acts upon an artificial, digital form of life called Avida, in which evolving computer programs compete for processor time. I study the monomer profiles in Avida that are loosely analogous to amino acid and carboxylic acid profiles in the terrestrial biosphere.

Chapter 3 extends the work in Avida with repeated evolutionary studies designed to examine the dependence of the monomer abundance profiles on environmental factors like mutation rate, initial conditions such as the ancestral genome, and alterations to the underlying physics of the system. The primary features of the MADB in Avida as measured in Chapter 2 are found to be largely independent of all other variables, indicating that selection is in fact the criterion that dominates evolved monomer profiles in this system.

The study examines each individual monomer's tendency to be selected for or against by altering its availability and measuring how that affects its representation in the ultimate evolved profile.

Chapter 4 continues the line of inquiry from Chapter 3 that investigated the dependence of individual Avida monomers' tendencies to be selected for or against by varying their frequency of appearance in mutation. In the further results given in Chapter 4, a much more detailed analysis varies the mutation frequency for each monomer in a systematic manner as opposed to the Monte-Carlo approach used in the prior work.

Chapter 2

Monomer Abundance Distribution Patterns as a Universal Biosignature: Examples from Terrestrial and Artificial Life

As revised for and resubmitted to *Journal of Molecular Evolution*, January 30, 2005.

Authors as published: Evan D. Dorn, Kenneth H. Nealson, Christoph Adami

2.1 Abstract

Organisms leave a distinctive chemical signature in their environment because they synthesize those molecules that maximize their fitness. As a result, the relative concentrations of related chemical monomers in life-bearing environmental samples reflect, in part, those compounds' *adaptive utility.* In contrast, rates of molecular synthesis in a lifeless environment are dictated by reaction kinetics and thermodynamics, so concentrations of related monomers in abiotic samples tend to exhibit specific patterns dominated by small, easily formed, low-formation-energy molecules. We contend that this distinction can serve as a universal biosignature: the measurement of chemical concentration ratios that belie formation kinetics or equilibrium thermodynamics indicates the likely presence of life. We explore the features of this biosignature as observed in amino acids and carboxylic acids, using published data from numerous studies of terrestrial sediments, spark-synthesis experiments, and meteorite bodies. We then compare these data to the results of experimental studies of an evolving *digital life* system. We observe the robust and repeatable evolution of an analogous biosignature in a digital life form, suggesting that evolutionary selection necessarily constrains organism composition and that the monomer abundance biosignature phenomenon is universal to evolved biosystems.

2.2 Introduction

The environmental concentrations of related chemical species carry information about their origin and synthesis. Relative rates of synthesis for individual chemical species differ between biotic and abiotic environments. Where molecules are synthesized by abiotic processes, rates of formation are constrained by the laws of thermodymamics and kinetics, resulting in a distribution of molecules dominated by low-weight and kinetically allowable species. Organisms, on the other hand, contain machinery (e.g., in terrestrial biota, enzymes) and expend energy to synthesize specifically those molecules they need for survival and competition. In the presence of life, therefore, some specific complex and high-formation-energy molecules are synthesized rapidly because they convey a fitness benefit. In effect, evolution acts on environmental chemistry by selecting for genomes that synthesize molecules at adaptive rates. If the compounds are stable (e.g., protected from rapid UV photolysis), then biosynthesis may be a significant factor governing the observed concentrations in a sample. Similarly, biologically mediated diagenesis may affect environmental concentrations in a way that reflects evolved metabolic activity and therefore, selection.

The distribution of a set of molecules in an environmental sample, therefore, may indicate the presence or absence of life. We call this effect the "monomer abundance distribution biosignature" (MAD Biosignature or MADB) and hypothesize that it is universal to all forms of life and collections of monomers they employ. While in principle, we may observe such a biosignature in any set of chemical measurements, in practice it is easiest to demonstrate by comparing the relative concentrations of chemicals within a single family of related molecules (such as amino acids) where comparison of synthesis rates is arguably valid. If our hypothesis is true, it carries implications both for the understanding of evolving biochemistries and the detection of extraterrestrial life, where the biochemistry of a putative ecosphere is not *a priori* known.

While we apply the MAD Biosignature to the challenge of detecting life in an extraterrestrial environmental sample, the use of monomer distributions as a diagnostic tool for other purposes is well-established. Fatty acid and phospholipid distributions have long been employed in the study of microbial diversity and other fields of biogeochemistry (see e.g., [92,98,88]), and amino acid profiles are frequently used as indicators of the diagenetic state of organic matter [43, 28, 17]. Amino acid profiles have also been used to distinguish between the synthesis pathways underlying abiotic amino acid formation in carbonaceous chondrites [12,55]. Nucleic acid frequencies in genomes are highly variable, but have notable statistical properties that can distinguish kingdoms and smaller families [78,40], and patterns of nucleic acids have been shown to reflect selective constraints such as RNA folding, thereby distinguishing functional from non-functional RNA [79], and the likely availability of particular amino acids in the prebiotic environment [13]. Geochemical profiles have even been employed previously to detemine whether hydrocarbon compounds in Earth's mantle are of a biotic or abiotic source, through comparison to the contents of meteorites and known abiotic syntheses [87].

Amino and carboxylic acids are familiar chemical families in which terrestrial biochemistry exhibits a clear signature of life's presence, and we collect and compare here data from a number of literature sources to demonstrate the effect. The statistical significance of the monomer abundance biosignature in amino acids has been previously studied [33], and we review here the most salient results. In this study, we conduct a more thorough analysis of amino acids and include saturated monocarboxylic acids in order to demonstrate that the MADB phenomenon as observed in the terrestrial biosphere is not limited to only one monomer family.

Any proposed biosignature should be tested not only against terrestrial, but also extraterrestrial life. To quote Maynard Smith [84]: "So far, we have been able to study only one evolving system and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look at artificial ones." We would like to go a little further to suggest that, in order to *detect* extraterrestrial life, we should have an understanding of the fundamental dynamics of living and evolving systems, independent of the organisms' particular substrate (i.e., chemistry). *Digital life* [2,73] is an artificial form of life that provides just such an instance, and it has been used successfully as an experimental platform for the study of evolutionary dynamics [2, 3, 9, 15, 53, 54, 69, 70, 93, 94, 96]. In a digital life system, self-replicating computer programs evolve within a digital ecosystem, enabling precise and controlled studies of evolutionary dynamics.

The Avida Digital Life Platform [71] is ideal for our purposes, since it is a controlled and understood system unrelated to terrestrial life. A digital system also makes possible repeated experimentation with evolution. An introduction to digital life can be found in [2] while Wilke and Adami [93] should be consulted for a recent overview of research conducted with this form of life.

We examined the relative distributions of computer instructions (the fundamental "monomers" in Avida) of numerous evolved and evolving populations to characterize the nature and robustness of the abundance patterns they form in response to evolutionary pressures. Avida organisms ("avidians") consist of a single genome: a string of computer instructions. These genomes evolve as a result of externally imposed mutations and selection pressures (the environment). The computer instructions that compose avidians are weakly analogous to the amino acids, fatty acids, or other compounds that compose familiar biota in the sense that they are selectable, meaning only those instructions that carry a fitness advantange will be reproduced in future generations. Therefore, any general fitness advantage or adaptive utility of a single instruction should be visible in its bulk concentration in an evolved population.

2.3 Materials and methods

2.3.1 Terrestrial biochemistry

2.3.1.1 Data sources

We compiled a database of measurements of related monomer concentrations (131 samples of amino acids and 31 samples of straight-chain monocarboxylic acids) from a variety of publications including studies of formation synthesis, terrestrial water columns and sediments, and meteorites.

Our set of abiotic amino acid data included thirty measurements in total, fifteen from tholin and spark-synthesis experiments [47,59,57,63,74,77], one from a report of UV photolysis in deep-space conditions [64], one of amino acid synthesis in a proton-irradiatiated mixture [89], seven from the Murchison meteorite [20,22,23,27,35,36,37,50], and six from two other carbonaceous chondrites [27,83]. Based on composition and racemic mixture, the amino acids from these meteorites are generally agreed to be of abiotic origin [19,35].

The complementary data set of amino acids from the terrestrial biosphere included 125 measurements of extracts from a variety of terrestrial sediments and water columns [17, 18, 28, 42, 44, 46, 45, 48]. Our database of carboxylic acids from abiotic sources includes two measurements from spark synthesis studies [80, 97] and three from meteorite sources [82, 66, 52]. All five sources report only the concentrations of short-chain carboxylic acids (C12:0 and shorter).

We rejected a few otherwise interesting measurements of amino acids including simulated Martian conditions [1], early spark-synthesis studies [60], and amber-encased insects [91] because they reported measurements of too few amino acids.

We compiled two different data sets of carboxylic acid measurements from terrestrial sources. We include only the straight-chain saturated monocarboxylic acids so that the data can be compared to sources of data on abiotic formation that report only this subfamily of compounds. The first data set includes twelve measurements of the same low-weight acids as the abiotic data set, all taken from coastal marine sediments [81]. A second group contains twenty-four measurements of longer chain carboxylic acids C14:0 through C28:0 in a variety of sediments and soils [98, 90, 16, 5].

2.3.1.2 Data treatment

To combine the disparate sources of data into unified sets for comparison, we applied the following protocol:

1. All data were converted to molar units.
2. Residues reported as “coeluted” (e.g., glycine and glutamate reported together as GLX) were both given the reported value.
3. “Trace” values were given a concentration of 0.
4. Where enantiomers were reported separately, their sum was used in a combined column.
5. Values reported as “approximate” or “estimated” were used unchanged.
6. Where values were reported as “$<$ X”, we used X.

In addition, we applied the following further steps to the amino acid and long-chain carboxylic acid data sets:

7. Where a clear statement could be found in the text that a search was exhaustive (e.g., "all other residues were absent or present only at trace levels"), unreported measurements were assigned a zero.

8. Other unreported values were left blank and not included in averages.

9. Columns with more than 50% blank values in any one category were eliminated from the analysis (e.g., Sarcosine is quantified in many spark-synthesis and meteorite studies, but never in sediment studies).

The reduced amino acid database included eleven residues (from an initial thirty) that were quantified with sufficient frequency in both the biotic and abiotic categories: gly, ala, β-ala, aba, ser, ile, leu, asp, glu, thr, and val. The reduced long-chain carboxylic acid data set included all straight-chain acids from C9:0 through C28:0, except C27:0. The abiotic component of the short-chain carboxylic acid data set did not contain sufficient overlapping data for reasonable presentation of averages, so we have presented each measurement individually.

Each data set was then normalized to a value appropriate for that category. All amino acid measurements were normalized to the mole concentration of glycine in that sample. The short- and long-chain carboxylic acid data sets were normalized to the concentrations of propionic acid (C3:0) and palmitic acid (C16:0), respectively.

2.3.2 Digital life

2.3.2.1 The Avida environment

In the digital world of Avida, computer programs of varying length encoded in a simple language replicate themselves and compete for processor time and physical space. Avidians evolve by mutation and selection of a self-replicating genome with access to a "chemistry" of simple different instructions (the monomers; 29 were available in the variant used for this study). Typically, a population is seeded with a primitive organism made from only 13 of those instructions. As the organism replicates, mutations (including copy errors, spontaneous point mutations, and insert and delete mutations) are imposed randomly on all organisms in the environment at rates configurable by the experimenter. Mutations and insertions create a new instruction (randomly chosen from the library of 29 with equal probability for each) at the target location.

In order to replicate, an avidian needs energy in the form of CPU (central processing unit) time. Each organism can increase its fraction of CPU time by performing calculations on a stream of random input numbers. These calculations (computational tasks) are the analogue of metabolic reactions (in biochemistry), and the code evolved to perform these tasks is analogous to the genetic code for metabolic enzymes. The simple self-replicator used as a progenitor (ancestor) organism does not include code to complete any mathematical tasks. As genotypes evolve computational genes, they receive a higher proportion of CPU time and quickly come to dominate the population.

2.3.2.2 Evolved monomer abundances in terminal populations

Our first experiment examined the final distribution of instructions in 350 evolved populations that varied in physical parameters and initital conditions. In each run, we populated a grid of 3600 cells with one of two hand-coded ancestor genotypes, and allowed the population to evolve for 4000 generations. We quantified the bulk frequency of each of the instructions in the population every ten generations.

We used two different ancestor genotypes (carefully written to differ as much as possible from each other) to control for the effect of initial conditions on the final evolved biosignature. The two progenitors were Avida's default 13-instruction ancestor, which uses the instructions SEARCH-F, JUMP-B and INC for the primary functions of flow control and replication, and an alternate 55-instruction ancestor that replaces those instructions with CALL, RETURN, SEARCH-B, SHIFT-L, DEC, and JUMP-F. Both organisms are shown in Table 2.4.2.1.

We also varied the copy-mutation rate in seven levels ranging by factors of two from 0.00125 to 0.08 mutations per copied instruction. At each combination of ancestor and mutation rate, we performed twenty-five runs that differed only in random seed.

2.3.2.3 Time-evolution of a biosignature

We conducted a second experiment in Avida to show the time-course of the evolving MADB as a formerly abiotic environment becomes dominated by incident life. We generated an abiotic population by initializing the same 3600-cell Avida grid with randomly generated, nonviable (meaning nonreplicating) genomes. We created a lethal environment by bombarding the entire population with a high rate of point mutations. (A lethal mutation rate is so high that information cannot be maintained in a self-replicating genome; eventually the gene for self-replication is lost.) For the duration of the experiment, we periodically seeded this population with viable (mutation-free) progenitor organisms. This is analogous to a steady influx of space-borne spores onto a planet with a lethal environment and no existing biology. Then, we progressively stepped down the rate of point mutations until it reached a level where the organisms could survive and a population began to replicate and evolve. The population was allowed to evolve until the distribution of instructions had stabilized for several hundred generations. Then, the point mutation rate was stepped back up until the population died and the randomizing effect of the mutations returned the instruction distribution to its "prebiotic" state.

2.3.2.4 Data treatment

Of the twenty-nine instructions in Avida's library, we exclude one instruction, NOP-A, from the analysis. NOP-A is used by the system to initialize empty memory in dividing organisms and is therefore greatly over-represented in comparison to the other instructions, and fluctuates significantly as organisms execute the ALLOCATE instruction. Furthermore, NOP-A's base rate of synthesis and availability is strongly genotype-dependent. It cannot be estimated *a priori* and we cannot make reasonable statements as to whether its abundance reflects biotic or abiotic processes, so we therefore compare only the remaining 28 instructions.

We quantified terminal abundances in evolved populations by averaging the last ten measurements to provide data smoothing. Avidians in the process of replicating, and occasional nonviable mutations, cause small instantaneous fluctuations in population abundance of the individual instructions.

2.4 Results

2.4.1 Terrestrial biochemistry

2.4.1.1 Amino acids

Glycine and alanine, which have low molecular weight and whose synthesis pathways are kinetically more favorable, dominate mixtures of amino acids synthesized by abiotic processes. Some small non-protein amino acids, including sarcosine and beta-alanine, have been observed in synthesis experiments [47, 59, 57, 61, 63, 64, 77]. Other amino acids are present only in trace concentrations, if at all [59, 57, 61, 64]. This pattern reflects the thermodynamics and kinetics of amino acid synthesis and is remarkably consistent, regardless of the specific nature of the synthesis environment. For example, the alpha amino acids are likely formed via the Strecker (cyanohydrin) synthesis [61], which presumes the existence of precursor molecules that include the sidechain. The formation of heavier amino acids would require the prior formation of the larger sidechains: since these reactions would include their own kinetic barriers, the availability of the larger precursors, and ultimately, the final yields of the larger amino acids are expected to be lower. Furthermore, the larger amino acids generally have higher energies of formation [4] and therefore would be expected to have low equilibrium concentrations even in the absence of kinetic barriers.

Amino acids in terrestrial samples show a more varied distribution dominated by the protein amino acids in roughly equal proportions. Figure 2.1 shows the relative abundances of a set of twelve amino acids measured in a variety of biotic and abiotic samples. The amino acids are plotted in ascending order of ΔG_r (Gibbs free energy of synthesis) at 18^{o}C as reported by Amend and Shock [4] to show both the trend among abiotic synthesis products towards smaller and "simpler" compounds and the biosynthesis of complex, expensive compounds. The "Sediment" curve represents terrestrial sediment and water column extractions. "Meteorite" represents extractions from three carbonaceous chondrites, and "Synthesis" is an average of fifteen spark-synthesis experiments in a variety of atmospheres and a single extraction from amino acids synthesized via UV photolysis on analogues of interstellar ice. The Synthesis curve serves as an abiotic baseline: the pattern of amino acid concentrations we expect when life is not present. Except for a single anomaly (high glutamate concentration), the meteorite curve exhibits the same pattern, while the biologically generated Sediment curve differs significantly, exhibiting high concentrations of a number of the larger amino acids. The measured concentrations of the heavier protein amino acids vary greatly in biotic samples, but are consistently higher than is ever observed in abiotic sources.

This analysis only shows a small selection out of hundreds of possible amino acids. A quantification of all possible low-molecular-weight amino acids would show several such as sarcosine that are non-zero in the abiotic baseline but low or zero in the biotic curves, similar to beta-alanine in Figure 2.1. Many heavier amino acids would not be present in either case. Therefore, the complete amino acid biosignature for terrestrial organisms would appear graphically as twenty peaks for the protein amino acids and a few dozen smaller peaks for non-protein amino acids used for other biological functions such as synthesis intermediates, neurotransmitters, and breakdown products, in a large spectrum where most other residues were absent. Fig. 2.1 is a subset of this hypothetical biosignature plot.

2.4.1.2 Carboxylic acids

Preparations from spark-synthesis studies exhibit carboxylic acids of low carbon number: only acids C6:0 and smaller are seen in significant quantities, and acids larger than C12:0 are not found at all [80, 97]. Figure 2.2 shows the relative

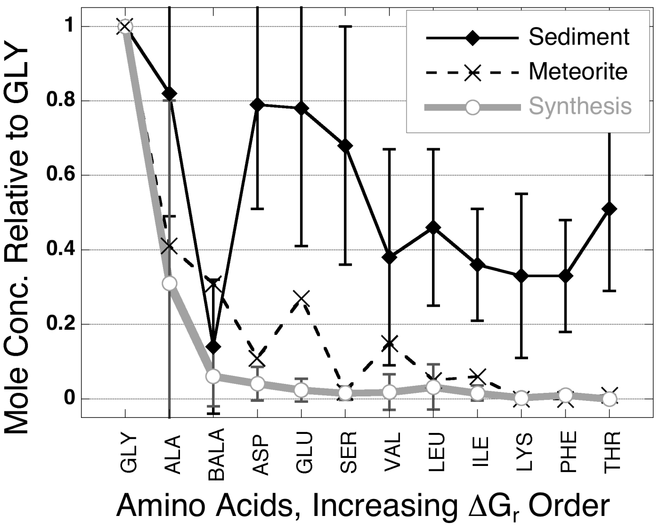


Figure 2.1: Average patterns of amino acid abundances relative to glycine, compared between biotic (Sediment n=125) and abiotic (Meteorite n=15, Synthesis n=16) sources. Error bars are one standard deviation.

concentrations of the first few acids measured in two such experiments and as measured in three meteorites, compared to an average of twelve measurements of terrestrial sediments. We observe in the Synthesis data a consistent exponential decrease in concentration with increasing molecular weight, while the Sediment curve shows a much more even distribution. The meteorite data fall between the two extremes, with a shallower exponential decrease and some variability in the relative concentration of acetic acid, probably due to its relatively high volatility and possible outgassing in space or during atmospheric entry. We contend that the pattern reflects the same fundamental processes that are at work in the amino acid data. As molecules become larger, formation energies are higher, synthesis kinetics generally become less favorable, and the number of possible conformations is greater, reducing the expected relative concentration of any particular conformer. The data from the meteorites implies a probable abiotic origin with a small amount of contamination by terrestrial material. We are not aware of any significant quantities of carboxylic acids larger than C12:0 recovered from non-contaminated meteorites. Long-chain (up to C33) lipid compounds have been created in abiotic synthesis at high (200 C and greater) temperatures, but the authors reported no bias toward any particular carbon

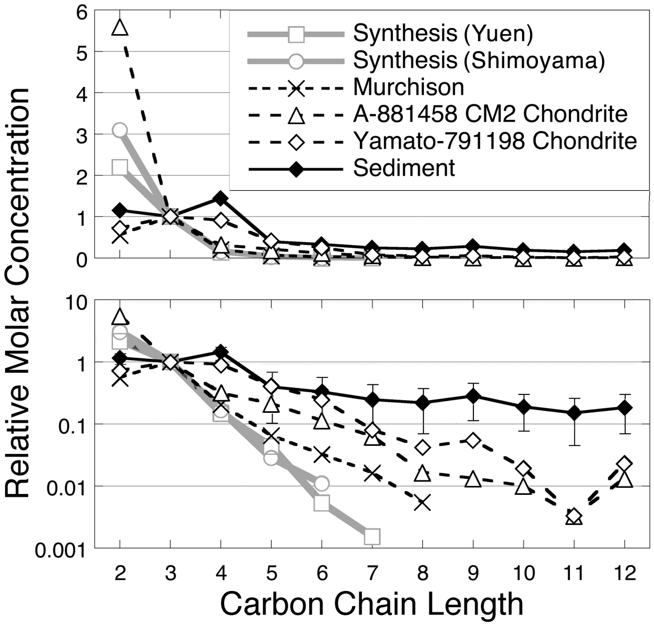


Figure 2.2: Relative concentrations of low-carbon number monocarboxylic acids from five abiotic sources compared to the average of 12 measurements in sediments. The data are plotted again on a semilog scale below to show the trend of decreasing concentration after C6:0 in abiotic samples. Synthesis curves represent two individual experiments (Shimoyama et al. 1994; Yuen et al. 1991).

number [75].

In terrestrial samples, however, the higher molecular weight acids C14:0 through C32:0 are present in abundance. As shown in Figure 2.3, environmental samples exhibit a consistent pattern that includes a primary peak at C16:0 with smaller peaks favoring even-carbon-number acids. This pattern cannot be explained by formation thermodynamics, but rather, reflects a pair of biological constraints. A fitness criterion is present: carbon chain lengths near 16 — in both single- and double-chain amphiphiles — are ideal for the formation of stable, non-porous, and flexible bilayers in aqueous solution at biologically relevant ranges of temperature and pH [41, 31, 38], and are therefore critical for membrane structure in a water-based biosphere. The predominance of even-numbered carbon chains reflects processes specific to terrestrial biochemistry: the primary fatty acid biosynthesis cycle terminates at C16:0 (palmitate), from which other fatty acids are synthesized through lengthening or shortening cycles that operate in two-carbon steps. A bias

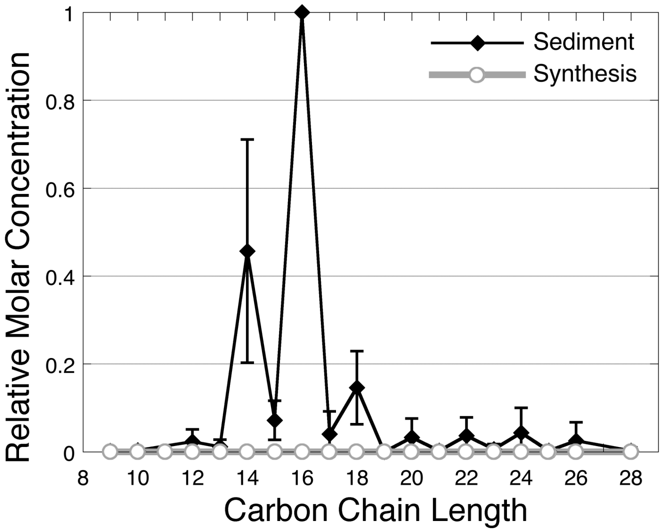


Figure 2.3: Curve representing the average pattern of long-chain monocarboxylic acids found in terrestrial sediments (n=26), compared to an abiotic baseline. Sediment data were normalized to C16:0 before averaging. Acids larger than C8:0 are not seen in measurable quantity in abiotic syntheses and so are shown as zero.

toward even-numbered carbon chains is considered diagnostic of biosynthesis and has been used to identify terrestrial contamination in meteorite sources [65]. The slight increase in concentration of C12:0 seen in two of the meteorites in Fig. 2.2 probably represents low-level terrestrial contamination.

2.4.2 Digital life

2.4.2.1 Monomer abundances in terminal populations

In Avida, two processes are present that we may consider "abiotic." First, we initalize the population with a hand-written, progenitor genotype. Second, mutations in Avida result in the replacement of an instruction with a new one, randomly chosen from the library with equal probability for each possible instruction. If selection did not affect the distribution of instructions, we would expect one or both of these effects to constrain the evolved abundance. If the initial conditions imposed by the ancestor constrained the evolved population, we would expect the final abundance to reflect the ancestor's distribution. If mutation were the dominant effect, as we would expect over long periods of time, the final population would show each of the instructions in equal proportion, demonstrating that the genotypes had incorporated the instructions at the same rate at which they appreared in mutation. Thus, the abiotic expectation in Avida, analogous to the Synthesis curves of Figs. 2.1, 2.2 and 2.3, is a flat line.

We observe in nearly all Avida runs, regardless of initial conditions or variables, a specific and consistent pattern of evolved instruction abundance that reflects neither of the abiotic constraints as a dominant feature. Figures 2.4a and 2.4b compare bulk instruction frequency curves to the abiotic expectation for populations from the two progenitor organisms in Table 2.4.2.1. Experiments were conducted at a number of different mutation rates (μ represents the probability of mutation at each genome site when the organism divides). While some variability is present as conditions change, many of the general features of the pattern are conserved and all evolved patterns are distinguishable from the abiotic baseline. The relative abundances observed in our experiments differ not only in their mean, but also in their variance. Instructions critical to an organism's survival (such as COPY and DIVIDE) are usually present at predictable levels (here, one per genome), and therefore show little variability. Optional instructions such as NAND, on the other hand, are expressed at high but variable concentrations as organisms in different competitive environments use them to complete different mathematical tasks (see "The Avida Environment" in Materials and Methods). Instructions that are frequently lethal when they appear as mutations (e.g., RETURN and JUMP-F) are suppressed, and the useless but non-fatal instruction NOP-X appears at a low but persistent level.

The distribution of relative abundances is essentially consistent over a large range of mutation rates (factor of 32), and changes only as the mutation rate becomes so high that the majority of offspring contain at least one mutation. At a near-lethal mutation rate of μ=0.08, a change in survival strategy generates a dramatic shift in the biosignature. Organisms abandon mathematical tasks and opt for very short genomes in order to maximize the likelihood of spawning genetically pure offspring. Mathematics-related instructions (NAND) are suppressed, while instructions that generally appear only once per genome (COPY, ALLOCATE, DIVIDE) are present at a higher concentration in the population as a whole because of the shorter average

Progenitor	**Alternate Progenitor**		**Evolved Avidian**	
1 SEARCH-F	1 CALL	29 DEC	1 SEARCH-F	29 NOP-C
2 NOP-A	2 NOP-A	30 NOP-C	2 NOP-A	30 PUT
3 NOP-A	3 NOP-A	37255	3 NOP-A	31 NOP-X
4 ADD	4 SHIFT-L	32 NOP-C	4 ADD	32 DIVIDE
5 INC	5 ALLOCATE	33 PUSH	5 ADD	33 PUT
6 ALLOCATE	6 PUSH	34 NOP-C	6 GET	34 SHIFT-R
7 PUSH	7 SWAP-STK	35 RETURN	7 ALLOCATE	35 DEC
8 NOP-B	8 PUSH	36 NOP-X	8 NAND	36 PUT
9 POP	9 NOP-B	37 NOP-X	9 PUT	37 NOP-B
10 NOP-C	10 POP	38 NOP-X	10 NOP-C	38 COPY
11 SUB	11 NOP-C	39 NOP-X	11 PUT	39 INC
12 NOP-B	12 POP	40 NOP-X	12 NAND	40 IF-N-EQU
13 COPY	13 NOP-B	41 NOP-X	13 GET	41 NOP-A
14 INC	14 IF-N-EQU	42 NOP-X	14 NAND	42 JUMP-B
15 IF-N-EQU	15 NOP-A	43 NOP-X	15 GET	43 NOP-A
16 JUMP-B	16 CALL	44 NOP-X	16 PUSH	44 GET
17 NOP-A	17 NOP-A	45 NOP-X	17 PUT	45 NOP-B
18 DIVIDE	18 NOP-B	46 NOP-X	18 POP	46 NOP-B
19 NOP-B	19 DIVIDE	47 NOP-B	19 NOP-X	47 JUMP-B
20 NOP-B	20 JUMP-F	48 NOP-B	20 IF-N-EQU	
	21 NOP-B	49 SEARCH-B	21 NAND	
	22 NOP-C	50 NOP-A	22 IF-BIT-1	
	23 COPY	51 NOP-B	23 NAND	
	24 INC	52 RETURN	24 GET	
	25 POP	53 NOP-X	25 NAND	
	26 NOP-C	54 NOP-C	26 NAND	
	27 DEC	55 NOP-A	27 NOP-C	
	28 NOP-C		28 PUT	

Table 2.1: Examples of three Avida organisms. The progenitors, used to seed initial populations, were hand-coded and are poorly adapted to the Avida environment. The "evolved avidian" is an example genome that arose approximately 1000 generations into a typical run seeded with the standard progenitor.

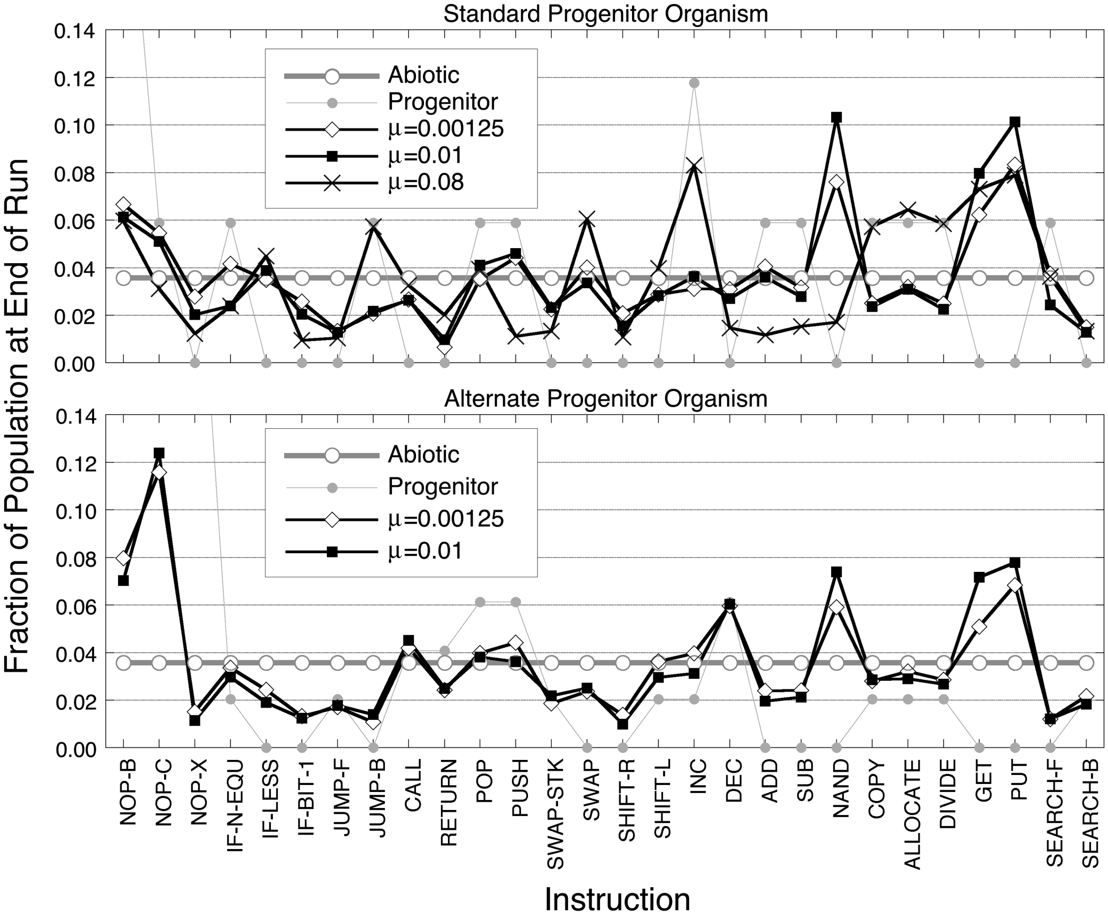


Figure 2.4: Relative abundance curves for several evolved populations. Note how a wide range of mutation rates (μ=0.00125 to 0.01) produces nearly identical distributions. Tests at numerous intermediate values of μ showed similar results (not shown). Near-lethal mutation rates (μ=0.08) and alternate progenitor organisms produce distinct distributions that share some features with the "standard" distribution. n=25 for each curve.

genome length. Other features, such as suppression of often-lethal mutations (JUMP-F, RETURN) are retained.

The observed pattern of evolved instruction abundance retains most of its salient features when we seed the population with a drastically different progenitor organism (See Table 2.4.2.1), demonstrating that the constraints of adaptation generally overwhelm the effect of initial conditions. As runs progress, the signatures of populations descending from the two disparate progenitors tend to converge. However, there are some persistent features that most likely reflect weakly connected adaptive peaks in the landscape. These local fitness maxima represent distinct genotypes that are fairly well adapted themselves, but cannot be derived from each other via evolution because the intermediate mutations required to convert one genotype into the other would be invariably fatal. For example, the two progenitors differ in implementing their essential "copy loop": one uses the INC (increment) operator while the other uses DEC (decrement). Changing INC to DEC in the copy loop would require several simultaneous mutations, any one of which would be fatal alone. As a result, this transformation is never seen in practice. Though INC and DEC may serve other functions in an organism, evolved populations tend to express the operator used by their ancestor's copy loop. This effect is evident in Fig. 2.4.

2.4.2.2 Time-evolution of a biosignature

When viable self-replicators begin to dominate a previously lifeless Avida grid, their signature frequency of instructions quickly overwhelms the pre-existing random distribution. Figure 2.5 shows the time-varying abundances of several instructions in one such run. At the outset, all instructions were present in equal abundance as the cells contained randomized, nonviable code — exactly the hypothetical abundance pattern used as an abiotic baseline in Fig. 2.4. When the point mutation rate was decreased to a non-lethal level, life immediately became the dominant process.

As the population filled the available space, the distribution of instructions in the ancestor's genome began to dominate, but was quickly replaced as higher-fitness organisms evolved. A recognizable distribution evolves with the same features we

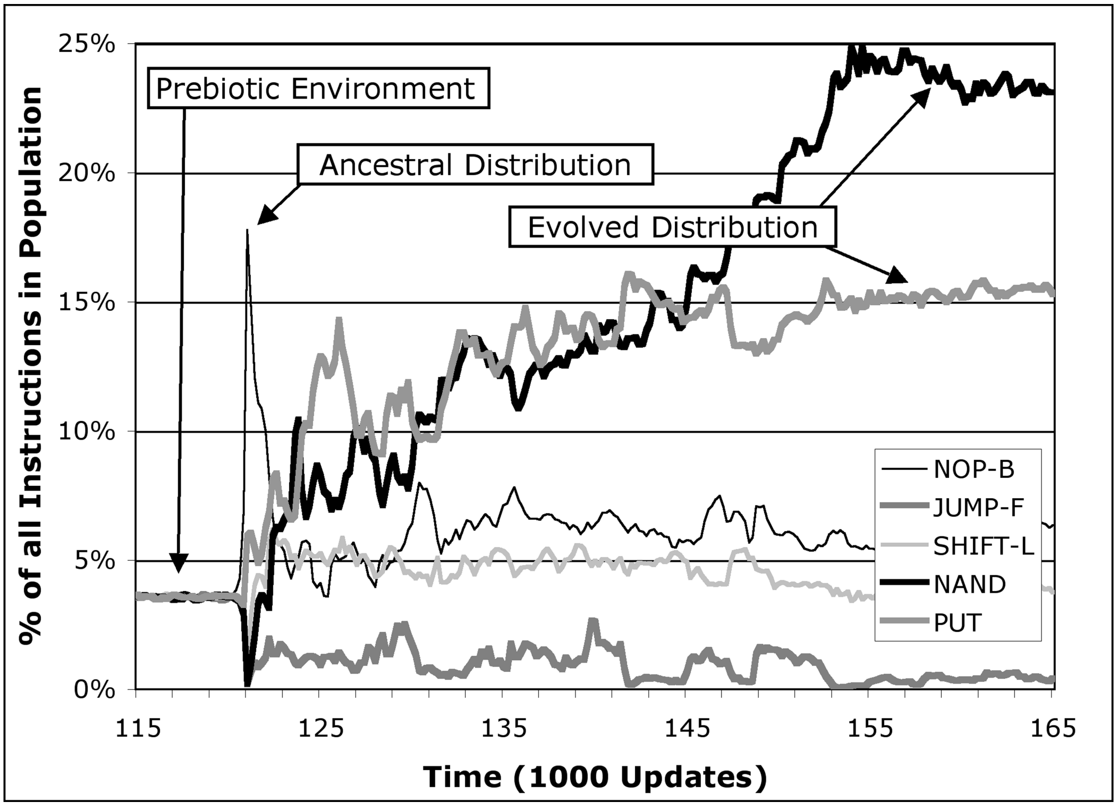


Figure 2.5: Abundance vs. time for 6 instructions as an incident life form populates a previously abiotic environment. The ancestor's most abundant instruction is NOP-B, while the adapted population's most abundant instruction is NAND. All instructions were present in equal proportion in the initial, mutation-dominated environment that did not contain any replicating organisms.

see in Fig. 2.4, such as high NAND concentration, but suppressed JUMP-F. The distribution stabilizes after a few hundred generations, even though the organisms themselves were still evolving.

We tracked all 28 instructions and allowed the population to evolve for several thousand generations beyond the point shown in Fig. 2.5. Eventually, we increased the mutation rate to its previous level, and the population died. At that point, the bombardment of point mutations rapidly returned the environment to its prebiotic state. Digital movie files showing the time evolution of all 28 instructions for the full time course of the experiment shown in Fig. 2.5. — and the development of one run used to generate Fig. 2.4 — may be found in the online supplemental information. (Note: these files have also been included in the electronic supplement to this thesis, see Appendix A)

2.5 Discussion

We believe the monomer abundance distribution biosignature is a universal feature of evolving systems: rapid synthesis of a select subset of compounds by a living metabolism should inevitably leave a biosphere in notable disequilibrium with respect to a sterile planet of the same geochemical composition. If we consider any particular monomer family, it is unlikely that thermodynamics would constrain abiotic synthesis to precisely the same rates as biotic synthesis. Indeed, if abiotic synthesis did produce monomers in the same concentrations as evolved biosynthesis, then biosynthesis would not be necessary, selection could not improve on abiotic geochemistry, and we question whether such a circumstance could in fact be described as "life."

Furthermore, the MADB should be detectable whenever life is present. While both abiotic and biotic syntheses may be present in an environment, we may reasonably expect biosynthesis to dominate most molecular synthesis as it does in the terrestrial biosphere. The coupling of energy from metabolism to catabolism, the manufacture of catalysts (enzymes), and the ability to concentrate reactants inside cells conspire to generate vast increases in synthesis rates relative to abiotic chemistry. As a result, even small quantities of life should leave a detectable chemical signature.

Understanding evolution's effects on environmental biochemistry carries significance, both for our general understanding of life and for the search for extraterrestrial life. It is essential for the latter because we cannot predict the specific biochemistry of the putative life we are searching for. A hypothetical extraterrestrial biochemistry might use a different set of amino acids, or use them in different abundances. But while such a pattern may not look like the terrestrial data in Fig. 2.1, it would also be unlikely to match the consistent pattern seen in abiotic syntheses. Likewise, extraterrestrial fatty acid biosynthesis could conceivably use cycles that operated in one- or three-carbon increments, resulting in a different distribution of peaks. If the organisms used ammonia, methane, or some other primary solvent rather than water, the adaptive peak for stable membranes might not be C:16. In either case, the unique signature of this exotic biochemistry would reflect the adaptive constraints of these organisms' biosphere and would probably be very unlike the "Synthesis" curves in Figure 2.2.

A contained digital system like Avida has several advantages for studying the MADB hypothesis. The course of evolution can be studied repeatedly to provide statistical significance to dynamics that usually only yield singular, history-dependent events. With terrestrial life, we can only examine one example of an evolved biochemistry: we cannot "start life over" on Earth and measure the concentrations of amino or carboxylic acids in each resulting biosphere, and therefore cannot prove that the particular concentration patterns we observe are the result of evolutionary constraint rather than happenstance. In Avida, we can repeat the process indefinitely to obtain statistics about how organisms adapt. In addition, we have complete control over initial conditions, allowing us to accurately characterize their impact on the result. Finally, Avida is completely unrelated to terrestrial biochemistry, giving us an opportunity to study a biosignature protocol in a non-Earth-centric setting.

2.6 Conclusions

The relative distribution of abundances of a set of monomers is a strong biosignature that shows promise for detecting even unusual or unknown biochemistries. The constraints of adaptation overwhelm the abiotic chemistry present in the environment as organisms expend energy to build complex, low-entropy structures. The strong similarity between laboratory synthesis and meteorite samples is a remarkable result since meteorite samples may have formed under conditions very different from laboratory syntheses and have aged under planetary and/or deep space conditions for thousands or millions of years before collection and analysis.

Using a digital life platform, we have demonstrated that the monomer abundance biosignature phenomenon is germane to any arbitrary living system, and have studied the robustness of the marker across a range of mutation rates and seed organisms. While an exhaustive analysis of initial conditions cannot be performed even for this simple form of life, the robustness of discrimination we see, given the variety of evolutionary paths and organisms produced in these experiments, bodes well for the universality and applicability of the monomer abundance biosignature.

It may be possible to look for instances of the monomer abundance biosignature simply by empowering a life-detection system with a few basic models (derived theoretically and/or empirically) of expected abiotic chemical abundances, and then examining samples for patterns that differ significantly. Strong deviations from thermodynamic/kinetic expectations, i.e., the appearance of unusual concentrations of any high formation energy, low probability compounds, could signal that a site or sample may contain life and is worthy of further investigation. This strategy poses significant challenges, but ultimately frees the search for extraterrestrial life from biases centered in our present understanding of terrestrial biochemisty.

Creating accurate models of expected conditions will be difficult, as unfamiliar local effects and boundary conditions (e.g., excessive concentrations of a certain mineral, or unusual temperatures) could change the chemical species present in a sample and cause a false positive. However, we may find some safe generalizations (for example, that phenylalanine and arginine will always be absent or low in the absence of life) that enable astrobiologists to build reliable experiments.

2.7 Acknowledgements

The research described in this work was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA), with support from the Director's Research and Development Fund (DRDF) and from the National Science Foundation under contract No. DEB-9981397. We thank Claus Wilke, Ronald V. Dorn III, and Diana Sherman for discussions.

Chapter 3

Monomer-Distribution Biosignatures in Evolving Digital Biota: Robustness With Respect to Physical Laws

As submitted to *Artificial Life*, October 20, 2004.

Authors as published: Evan D. Dorn, Christoph Adami

3.1 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 their environment that is anomalous, given the local chemistry. We observe this effect in the distribution of computer instructions used by an evolving population of digital organisms, and 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.

3.2 Introduction

Evolved bio-organisms impress a distinctive chemical signature on their environment because biota synthesize those compounds necessary for competition and replication. More generally, biochemical synthesis may be seen as the product of selection, as evolution shapes genomes so as to maximize their fitness. As a result, the chemical species that persist in the environment do not generally reflect the chemical species we would expect if life were not present.

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, and therefore, 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 [33, 51, 59, 60, 62, 64, 95].

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' marginal utility in fitness rather than, or in addition to, their thermodynamic cost. If the compounds persist in the environment, we expect bulk environmental concentrations of these monomers to also reflect fitness criteria whenever biosynthesis is present at significant levels. This effect, which we call the "monomer abundance distribution biosignature" (MADB), is very pronounced in the terrestrial biosphere, and is easily detectable by a number of mathematical techniques [34, 33].

Previously, we demonstrated [34] that the MADB is observable and repeatable in at least one artificial evolving system, the Avida Digital Life Platform (an introduction to Avida may be found in [71]). Avida organisms ("avidians") are small, self-replicating programs written in a simple but robust and evolvable language; 29 instructions are available in the variant used for this study. The instructions may be seen as weakly analogous to the monomers such as amino acids that compose familiar biota. 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. This is the 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, indicating that selection has dictated the monomer abundance pattern of the population. This pattern is largely consistent over many trials, even though the actual genomes evolved may not resemble each other at all. Moreover, the pattern

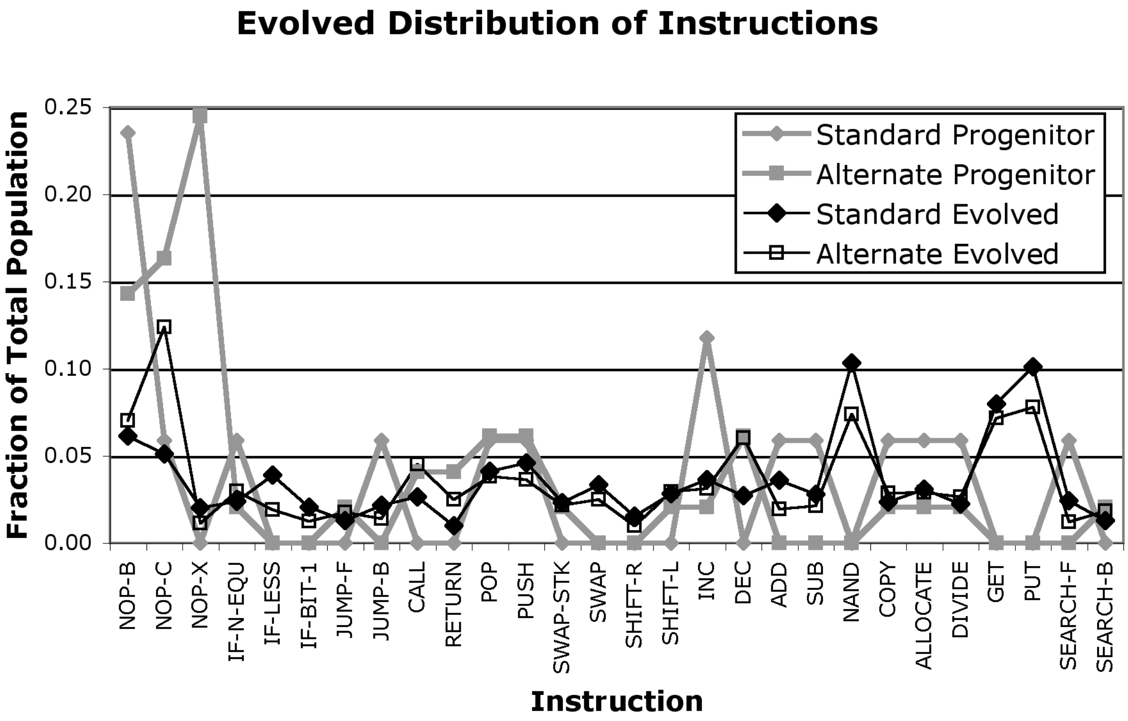


Figure 3.1: The distribution of instructions in two different ancestor organisms and in populations descended from those ancestors. While the ancestors have very different composition, the composition of the descendents is similar as the terminal populations have adapted to the same environment. "Evolved" lines represent the average of 25 different evolutionary trials, each sampled after 1500 generations.

is consistent over a wide array of parameters such as mutation rate and different ancestors. Figure 3.1 shows the relative distribution of 28 computer instructions in evolved Avida populations descending from two distinct ancestor genotypes. These results are more fully presented in [34]. Note that while the ancestors have very different composition, their descendents have converged to a common profile, demonstrating 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 Figure 3.2 (from [34]). In the experiment which produced Figure 3.2, 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

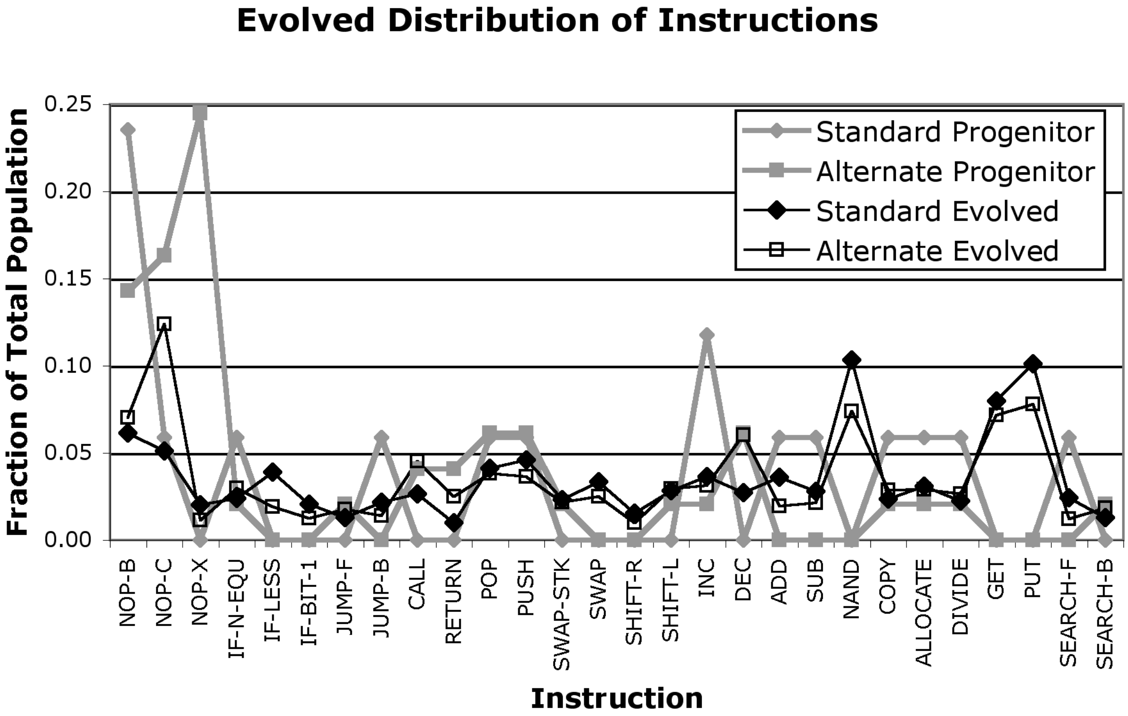


Figure 3.2: Figure 2. Evolution of the distribution of six computer instructions as incident self-replicators colonize and adapt to a formerly lifeless environment. At the outset, mutations cause all instructions to be present in roughly equal proportion. The ancestor organism's genome — nearly 20% NOP-B — dominates the early biotic distribution. As the organisms adapt to the environment, a NAND-heavy distribution develops and stabilizes. Often-lethal instructions such as JUMP-F are strongly suppressed by selection forces.

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 explore the robustness of the MADB as the fundamental abiotic parameters of the Avida environment are changed. This is important because it can be argued that the MADB observed in terrestrial biochemicals (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 distiction seen between biotic and abiotic patterns are not the product of selection, but of some other, unknown function. In artificial life, we can test these conjectures.

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, thus changing their availability to early life forms. 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' abundance will be more or less independent of the frequency with which they appear in mutation, indicating 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, thus ensuring 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.

3.3 Methods

By default, Avida substitutes new instructions during mutation events with an equal probability for each instruction. For this experiment we created a modified version of Avida that allows the experimenter to specify a probability of substitution for each instruction.

In each experiment, a grid of 3600 cells was populated with a 13-instruction simple self-replicating ancestor. This initial population was evolved for 1500 generations and the bulk frequency of each instruction in the population was quantified every 100 generations. To provide smoothing of momentary fluctuations in instruction concentration, we used the average of the last ten. We performed 25 replicates of each experiment.

We created eighteen different “spectra” that represented the relative frequency with which each instruction appeared in random mutation. Three of the spectra were manually constructed systematic variations where the instructions, listed in an arbitrary order (the default library order) were given increasing or decreasing representation such that the most-represented instructions were substituted 29 times as often as the least-represented instructions. Another fifteen experiments used randomly generated mutation frequencies.

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 to inititalize empty memory in dividing organisms. Therefore, the number of NOP-A instructions appearing in the population is highly genotype-dependent, but not in a way we can conclusively call “biotic” since the organisms do not have a choice about how empty memory is inititalized. It cannot be selected for or against in an evolutionary sense, and therefore we choose to exclude it from the analysis.

3.4 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 2 shows the average evolved abundances of the 28 instructions for four different mutational profiles. 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 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. Some other instructions, such as JUMP-F and RETURN, are frequently lethal when they appear in mutation. As a result, they are rarely incorporated into genomes and appear underrepresented in the final population regardless of their mutational frequency.

Figure 3.4 shows the relationships between mutation frequency and evolved population frequency for several instructions. In each of these plots, if selection did not constrain the organisms's use of each instruction, we would expect the data to fall on or near the unity line, reflecting that the organisms used the instructions at the same rate at which they appear in mutation. 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

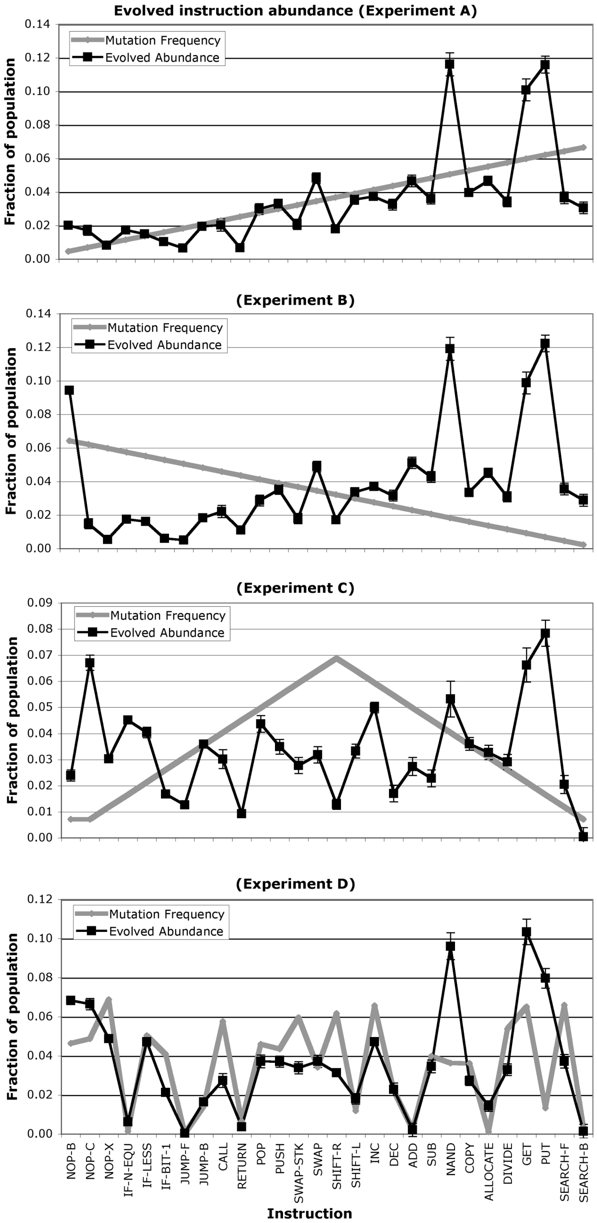


Figure 3.3: Four experiments showing the evolved distribution of instructions as the underlying physics are changed. The gray curves represent the frequencies with which each instruction was presented to organisms through mutation, black curves represent the relative abundances of the instructions in 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 similar randomly generated distributions tested.

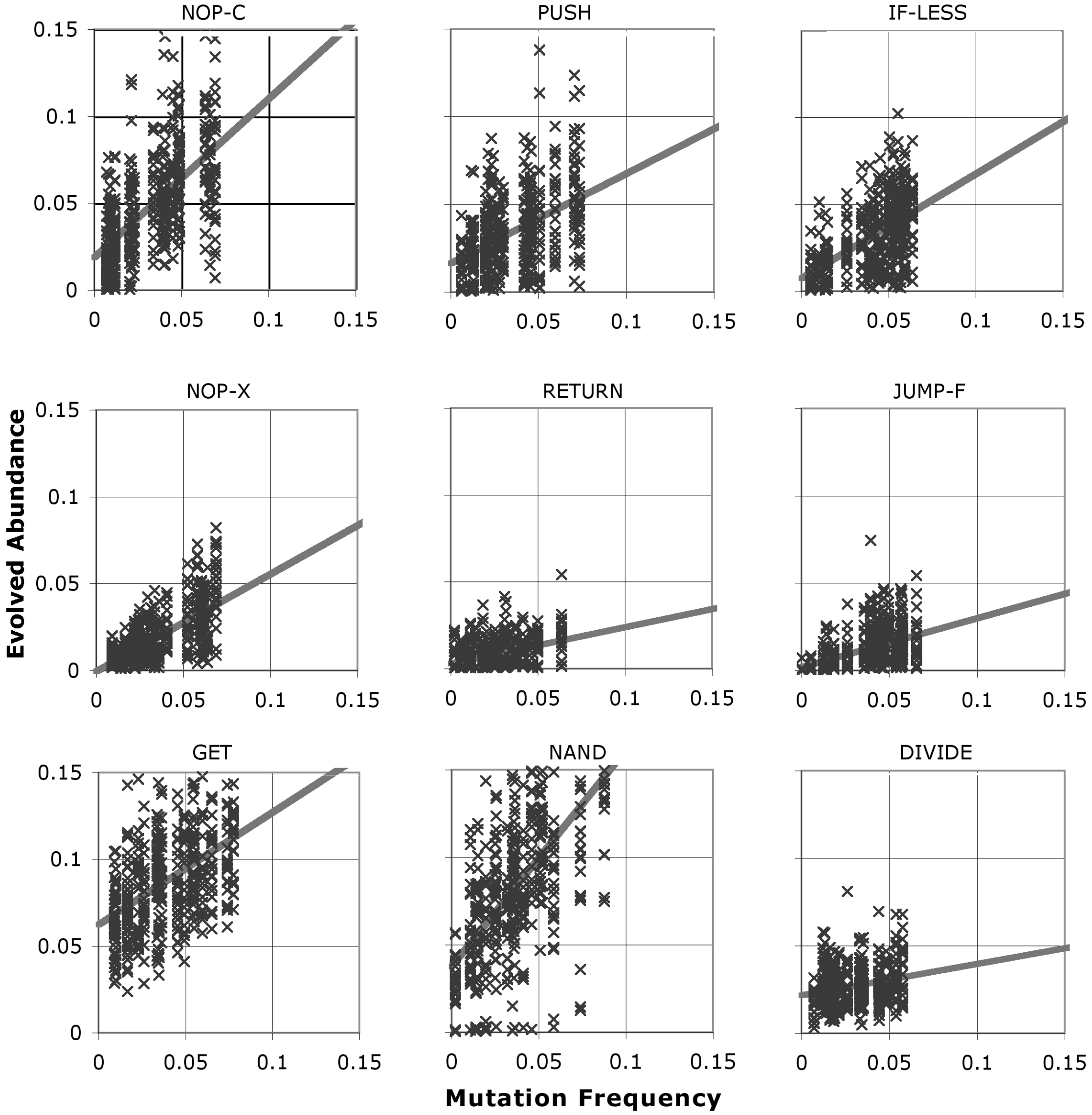


Figure 3.4: The relationship between frequency of appearance in mutation and abundance in the terminal population for nine instructions. GET typifies an instruction which frequently conveys a fitness advantage; it tends to be incorporated into genomes at a high level regardless of how often it appears as a mutation. Conversely, RETURN is not necessary for basic function and is often lethal as a mutation, so it remains at a low level, even when it appears frequently as a mutation. N=450, 25 runs each using 18 different mutation frequency profiles. Gray lines are a simple least-squares linear fit.

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 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 non-functional 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) are responsible for input and output within the computational metabolism; genomes cannot gain any advantage over the ancestor without using them, and 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. However, it shows a strong, in fact, greater than unity dependence on the input frequency of mutation. This almost certainly represents faster learning of tasks when NAND appears in mutation more often. Seventy-three tasks are rewarded in Avida, each of which requires one to five NAND operations. Each gives a small multiplicative fitness bonus, and previous experience has shown that avidians will continue to learn tasks for hundreds of thousands of generations until all possible tasks have been exploited. Since our experiment only tested 1500 generations, the high slope of the NAND distribution probably represents a strongly increasing tendency to learn more tasks in the time available as NAND appears in mutation with increasing 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 as 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 of 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 distribution seen in the plot is largely due to variation in the average length of the evolved genomes.

Table 3.4 gives full results for the selection bias of the 28 instructions we analyze. We use the summed deviation from the unity line over all data as an approximation of the tendency of selection pressures to elevate the population concentration of a single instruction; the instructions are listed sorted by this measure. Instructions with the strongest selection bias are at the top. Table 3.4 also shows the slope of the line fit to each instruction, which indicates the degree to which an instruction's population abundance depends on its mutation frequency. The least-dependent instructions (generally those which have lethal effect and are often suppressed) are listed first.

3.5 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,

Selection Bias		**Slope**	
PUT	31.1	DIVIDE	0.18
GET	21.6	RETURN	0.21
NAND	20.7	JUMP-B	0.25
NOP-B	12.6	JUMP-F	0.28
NOP-C	7.2	COPY	0.30
INC	1.9	SEARCH-B	0.33
SWAP	1.4	SEARCH-F	0.36
ALLOCATE	1.2	IF-BIT-1	0.36
ADD	0.2	CALL	0.39
PUSH	-0.4	SWAP-STK	0.45
POP	-1.1	INC	0.47
DIVIDE	-1.2	SHIFT-R	0.48
SHIFT-L	-1.6	PUSH	0.51
SEARCH-F	-2.6	SHIFT-L	0.55
IF-N-EQU	-3.0	ALLOCATE	0.55
JUMP-B	-3.4	NOP-X	0.56
COPY	-3.6	IF-N-EQU	0.57
IF-LESS	-3.9	IF-LESS	0.60
DEC	-4.3	DEC	0.62
SUB	-4.8	GET	0.65
CALL	-5.2	POP	0.65
SWAP-STK	-7.0	SUB	0.71
RETURN	-7.4	ADD	0.75
NOP-X	-7.6	PUT	0.83
IF-BIT-1	-9.1	SWAP	0.89
SEARCH-B	-10.2	NOP-C	0.91
SHIFT-R	-10.4	NOP-B	0.94
JUMP-F	-11.1	NAND	1.24

Table 3.1: Selection biases and the slope of evolved frequency vs. mutation frequency for 29 instructions.

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. Through a-life, 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 [14].

We have demonstrated the repeatability of the monomer abundance distribution biosignature in Avida populations, and characterized its robustness with respect to alterations in the underlying physics. We find 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. 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 conclusion indicates one possible direction for the search for nonterrestrial life via a method that is agnostic of terrestrial biochemistry, i.e., "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 of those ranges. Measurements of chemical concentrations that deviate from those ranges may indicate that an evolved metabolism is selectively synthesizing useful compounds.

This approach to life detection has obvious limitations and obstacles. 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. Any such experiment 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.

Nonetheless, we see the MADB as an important biosignature for life detection, worthy of investigation because it is the inevitable result of a fundamental life process (evolutionary selection) and 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.

Chapter 4

Monomer-Distribution Biosignatures in Evolving Digital Biota: Robustness With Respect to Physical Laws; Additional Data and Unpublished Results

The results shown in Figure 3.4 were produced by creating Avida runs in which the "availability" of each instruction was altered by changing the probability with which it was chosen when random instructions were selected for mutations. By increasing the probability with which a particular instruction was chosen, we increased the rate at which it was made available to be incorporated into avidian genomes.

In the experiment that produced Figure 3.4, fifteen profiles (spectra of availability) were created, each of which specified a probability of appearance for each instruction. These profiles were chosen randomly. Twenty-five runs were performed to 1500 generations for each profile, and the resulting monomer abundance profiles were quantified. (In addition, twenty five runs for each of the three systematically varied profiles shown in Figures 3.4a, 3.4b, and 3.4c were also included).

I was concerned that the Monte-Carlo approach might be contributing unnecessarily to the wide distribution and general noisy character of the data seen in Figure 3.4, because as the mutation frequency of one instruction changed, so did all the others, and interdependencies between the instructions might be altering the results. Also, the fact that NAND remained at or close to zero in some runs indicated that the avidians in those runs were not learning to perform tasks. However, since the avidians in most runs did learn — even at the same availability levels for NAND at which they sometimes failed — I was concerned that 1500 generations was possibly insufficient for adequate equilibration of the population.

To address these concerns, I performed an experiment in which each of the instructions was varied systematically: only one instruction was increased or decreased in mutation frequency relative to the others. This required a larger number of runs, as different runs were required for the characterization of each instruction. I ran ten populations for each of six different mutational frequencies for each of the 29 instructions. This experiment required approximately three months to complete using (on average) twenty-five Pentium-IV processors in a Beowulf supercomputing cluster.

Generally, the results were largely similar to those shown in Figure 3.4, and may be seen as confirming those results and refuting the notion that either the Monte-Carlo approach or the shorter run times of 1500 generations created any particular bias or contributed to the spread of the data. Figure 4.1 shows plots of the relationship between final population abundance and frequency of appearance in mutation for the same nine instructions depicted in Figure 3.4.

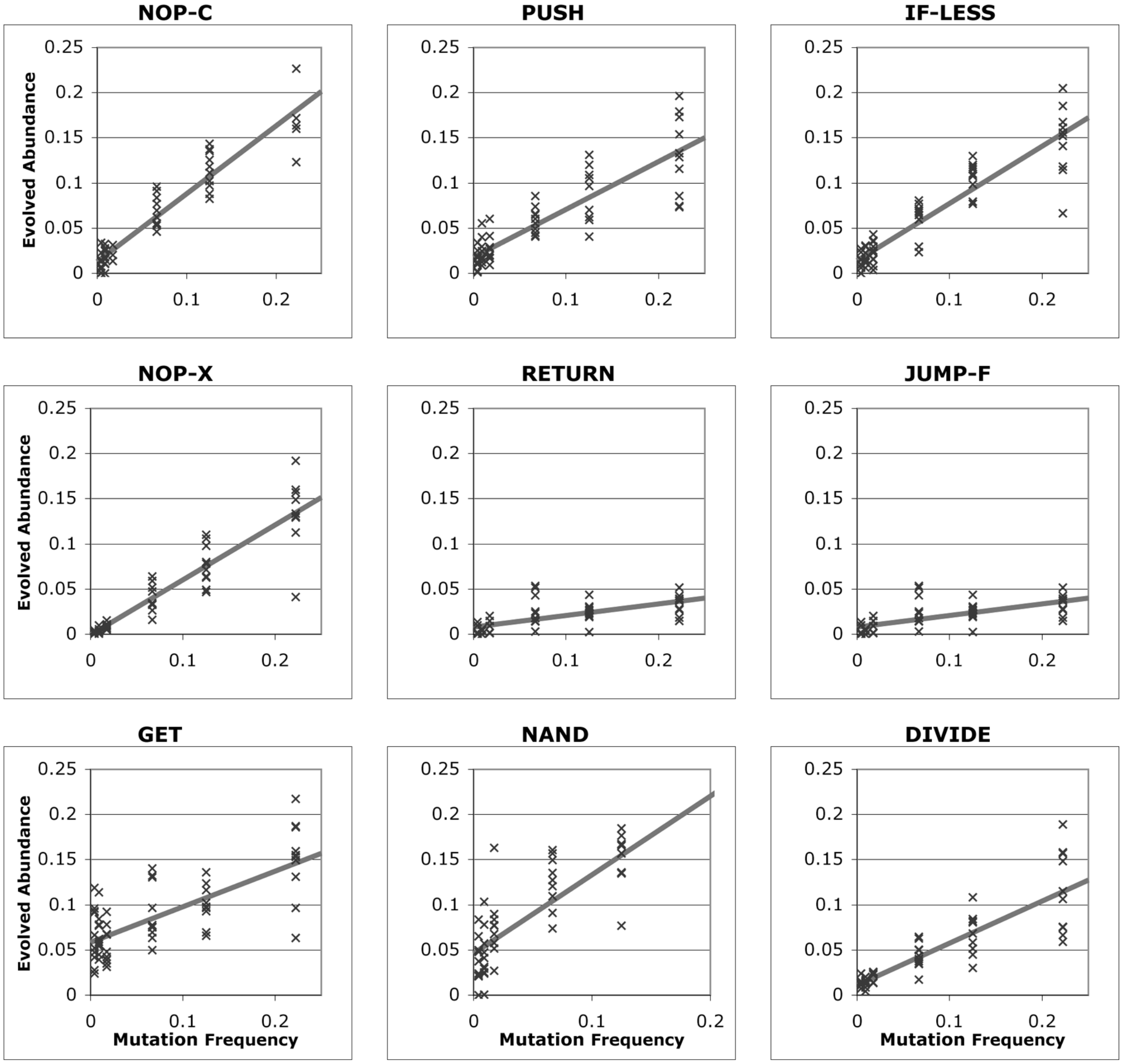


Figure 4.1: The relationship between final population abundance and mutational availability for nine instructions. Often-favorable instructions like GET and NAND generally appear at high levels, regardless of mutational ability because they are incorporated into genomes quickly. Often-fatal instructions like DIVIDE are heavily suppressed, maintained only at the minimum level necessary for replication. Neutral instructions like NOP-X tend to fall near the unity line, with an expression level in terminal populations that is highly dependent on their mutational availability.

Part III

Detection of the Monomer Abundance Distribution Biosignature

Part III comprises two chapters representing work related to automatic detection of the MADB with mathematical systems.

I examine the ability of pattern recognition systems, including principal components analysis and two varieties of neural networks, to detect the MADB as it appears in terrestrial samples of amino acids.

In Chapter 5 (published in *Icarus*, see [33]), I study the ability of numerical systems to differentiate between known examples of biotic and abiotic amino acid profiles, and explore the robustness of these systems with respect to noisy and partial data. The study goes on to examine the dependence of these pattern-recognition systems upon the individual elements of the monomer profiles by sequentially removing each amino acid from consideration and measuring the performance of the system when only the remaining data are available.

In Chapter 6 (under preparation for submission to *Astrobiology*) I consider a related problem geared specifically to the task of non-Earth-centric life detection: the ability of pattern recognition systems trained only on information about known abiotic patterns to correctly identify profiles that do not match the abiotic training set as exceptional and potentially biotic.

Chapter 5

Principal Component Analysis and Neural Networks for Detection of Amino Acid Biosignatures

As published in *Icarus*, 166 (2004), 403-409.

Authors as published: Evan D. Dorn, Gene D. McDonald, Michael C Storrie-Lombardi, Kenneth Nealson

5.1 Abstract

We examine the applicability of Principal Component Analysis (PCA) and Artificial Neural Network (ANN) methods of data analysis to biosignature detection. These techniques show promise in classifying and simplifying the representation of patterns of amino acids resulting from biological and non-biological syntheses. PCA correctly identifies glycine and alanine as the amino acids contributing the most information to the task of discriminating biotic and abiotic samples. Trained ANNs correctly classify between 86.1% and 99.5% of a large set of amino acid samples as biotic or abiotic. These and similar techniques are important in the design of automated data analysis systems for robotic missions to distant planetary bodies. Both techniques are robust with respect to noisy and incomplete data. Analysis of the performance of PCA and ANNs also lends insight into the localization of useful information within a particular data set, a feature that may be exploited in the selection of experiments for efficient mission design.

5.2 Introduction

We are studying how computer analysis techniques may be applied to detection of chemical biosignatures. Techniques such as Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) show promise for automated sample evaluation and as tools for the design of efficient experiments [58]. In this study, we explore the ability of these techniques to classify chemical biosignatures, specifically, measurements of amino acid frequency available in the literature. In coming years, we will have the opportunity to search for life on and below the surfaces of Mars, Europa, and other planetary bodies. Probes sent to these destinations will need to function with some degree of autonomy, as they will be subject to long communication delays and temporary communication blackouts [67]. Automated computer analysis of scientific data will be a critical element of all such missions. We believe that kinetic and thermodynamic constraints on the chemistry of living systems will result in chemical signatures that differ from the background of an organism's environment. If this statement is true of all living systems — earthlike or otherwise — systematic comparison of chemical samples to expected environmental norms can provide a non-Earth-centric method of life detection. We explore the ability of some techniques to distinguish samples of amino acids that have been synthesized by biotic and abiotic chemical systems. These techniques also hold promise for the improvement of mission design. By analyzing the contributions of different scientific measurements, we can determine the intrinsic value of each. This information is useful both for the selection of instruments and design of experiments.

5.3 Amino acid data set

We chose amino acids as the test case for this study because they have been previously studied as important biomarkers [7] and because substantial data already exist. To create a consistent data set for this study, we collected data on amino acid concentrations from four different sources. To represent environmental biomarkers, we used 79 measurements of amino acids extracted from marine sediments and water columns [18, 44, 28]. We make the assumption that, within the modern terrestrial biosphere, biological synthesis of organic molecules overwhelms any abiotic chemical synthesis present. To supplement the sediment data, we also included the average relative frequencies of amino acids in sequenced proteins from 106 protein superfamilies [29, 30]. These data differ from the others because they are computed from protein sequences rather than measured in extracts from environmental samples. As such, they do not reflect any bias that might be inherent in the extraction or analysis processes, or in potential differences in diagenesis. Furthermore, they carry their own intrinsic bias; sequences of individual proteins do not necessarily reflect the relative abundance of the amino acids in an intact organism or the community of organisms present in marine sediments. Nonetheless, we included the sequence data because they potentially help us gauge these biases. The combined sediment and protein data will be referred to as the "biotic" data set. To represent amino acids synthesized by non-biological processes, we employed the results of several spark synthesis experiments meant to simulate formation of organic molecules in environments resembling those of the early Earth [76] and other planetary bodies [57, 59]. In addition, we included in the "abiotic" data set measurements of amino acids present in the Murchison carbonaceous chondrite [24, 26, 36, 37, 35], and one report of amino acid formation on interstellar ice analogues [64]. Despite reports of small enantiomeric excesses of some amino acids in the Murchison meteorite [25], most scientists now agree that the amino acids are the result of non-biological synthesis and that contamination by terrestrial sources of organic molecules is low [26]. Our analysis will support that hypothesis. Each of the studies from which we used data sampled a number of different amino acids — up to a hundred in the case of some of the Murchison studies. We retained only the twelve amino acids which were most frequently measured in this set of studies: glycine, alanine, leucine, isoleucine, lysine, serine, threonine, aspartate, glutamate, phenylalanine, valine, and arginine. In some cases, certain of these amino acids were not looked for and/or not detected in a particular study. Wherever a measurement was unavailable, a zero was inserted in the combined data set. The data are represented as relative molar percentages normalized to the sum of the quantities of these twelve amino acids. Other amino acids, if present, are ignored entirely, as is the overall concentration of amino acids in the sample. Where both D- and L- enantiomers were quantified in the source, we used their sum. Figure 5.1 shows the average concentration of each amino acid by category. Note that the abiotic data contain high concentrations of glycine and alanine, providing an immediate selection criteria.

We also sometimes represented these data as the relative elemental (H/C, O/C, N/C) ratios present within the samples, computed from the measured relative amino acid concentrations. In this version of the data set, each sample had only three associated variables. It was useful to examine the data represented this way because measurements of element ratios are comparatively easy to make in the field. If we could perform these analyses using only the elemental ratios, a mission could be simplified. We were concerned that the different numbers of data points in the two

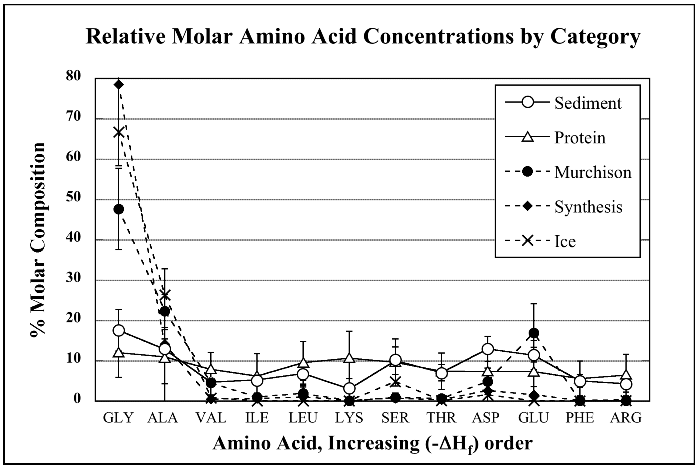


Figure 5.1: Amino acid concentrations represented in the data set. Each line is the average over all samples in that category (Sediment n=79, Protein n=106, Murchison n=7, Synthesis n=8, Ice n=1). Error bars represent one standard deviation.

categories (166 points from biological sources, 15 from non-biological sources) could introduce an unwanted bias in the analysis or make adequate discrimination difficult, as will be discussed in more detail below. In order to obviate this concern, the data set used for some operations was expanded with extra copies of the abiotic data until the total number of data points in each category was equivalent. This was performed several times with different amounts of random noise (from 0 to several times the standard deviation of the entire set) added to the extra copies. Results of the classifying algorithms showed little dependence on the amount of noise added.

5.4 Methods

5.4.1 Principal component analysis (PCA)

Principal Component Analysis, also known as the Hôtelling transform, is a linear analysis technique that finds the most efficient representation (in the least-squares sense) of a data set in several dimensions. It is often employed in data representation and data compression tasks, where representing a large data set in a smaller number of dimensions may be desirable. In general, with a set of M vectors $\boldsymbol{x}_m$, each with N elements such that $\boldsymbol{x}_{m,n}$ represents the nth element of the mth vector, PCA finds the linear combinations of the dimensions that encode the greatest proportion (in the least squares sense) of the variance in the data set $x_1 \ldots x_M$. PCA computes the zero-mean data set by subtracting the average vector from all data, and then finds the covariance matrix $\boldsymbol{C}$ where element $C_{i,j}$ of the matrix is defined as the expected value of the product of elements i and j in any individual vector $\boldsymbol{x}_m$:

$$C_{i,j} = \langle x_{m,i} \cdot x_{m,j} \rangle. \tag{5.1}$$

The normalized eigenvectors of this matrix, ranked by their corresponding eigenvalues, are the principal components of the data set. Projection onto the first p principal components is the most efficient linear representation of the data in p dimensions. Given a data set of moderate size, this algorithm is relatively robust to noise and is useful in its capacity to combine unrelated measurements into a common statistical framework [85]. By its nature, PCA normalizes all components of the input vector to unit variance, allowing us to combine measurements from different scales. PCA is best employed as a tool to reduce the dimensionality of a set of data. This reduction of dimensionality can allow one to visualize a multivariable data set more easily, and to employ traditional statistical methods that might otherwise be impossible to use. For example, a previous study in our laboratory used PCA to compare the amino acid patterns found in Murchison and ALH84001 with those found in proteins and Antarctic ice, concluding that the meteorite samples were similar to each other and to the ice, but that neither was similar to proteins [58].

5.4.2 Artificial neural networks (ANN)

Neural networks are among the simplest and best-understood pattern recognition algorithms available, and we will show that they are entirely adequate for the analysis of this data set. More robust (in terms of sensitivity to noise and incomplete data) and more versatile than linear techniques such as PCA, they present an ideal tool for biosignature detection and classification. An analog, feed-forward neural net is a

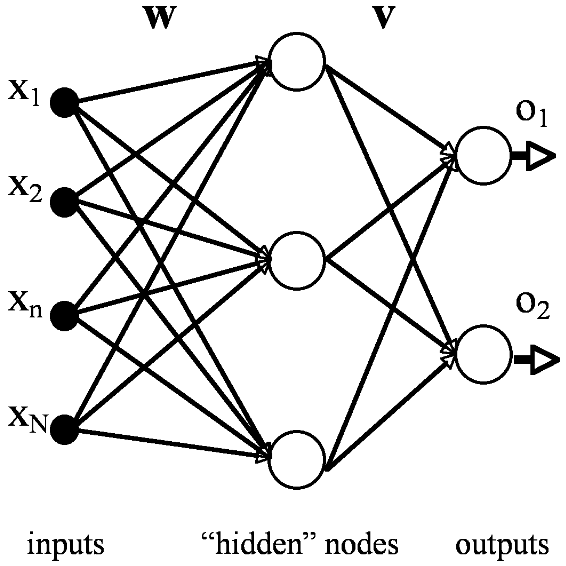


Figure 5.2: Schematic representation of a two-layer feed-forward neural network with input vector $\boldsymbol{x}$ and outputs O_1 and O_2

collection of "nodes", each of which implements the function:

$$O = \sigma(\sum_{i=1}^{n} x_i w_i). \tag{5.2}$$

A group of N inputs x is each multiplied by a "weight", the products summed and limited by a threshold function, typically an exponential or hyperbolic tangent sigmoid function that varies from 0 to 1. The two-layer network shown in Fig. 5.2 implements the function $O = \sigma(\boldsymbol{v}^{-1}\sigma(x\boldsymbol{w}))$ where $\boldsymbol{w}$ and $\boldsymbol{v}$ represent the first- and second-layer weight matrices. It can be proven that with a sufficient number of nodes, any mathematical function may be approximated with arbitrary accuracy through the selection of appropriate $\boldsymbol{w}$ and $\boldsymbol{v}$.

Computing the weights is called "training" the network and cannot usually be done deterministically. Many ANNs — including all of ours — are trained by the stochastic algorithm: the weight matrices are adjusted by random quantities $\boldsymbol{w}$ and $\boldsymbol{v}$ and the performance of the network is measured by computing the sum-squared-error over a set of training data. $\boldsymbol{w}$'s and $\boldsymbol{v}$'s which result in reduced errors are kept. Others are rejected except for a small percent of "wrong" answers, which are retained to allow the converging solution to escape local minima. In general, ANNs gain an advantage over PCA techniques because the transfer function may be nonlinear. A neural network can discover nonlinear interactions between variables in the data set that would be missed by a linear technique. In addition, sigmoidal σ functions saturate at large input values, helping to prevent extreme outlying data points from skewing the entire analysis. ANNs are generally very robust in nature. A well-trained network will usually make a correct analysis even when large amounts of noise have been added to the input vector and/or portions of the data are unavailable. In our case, simple neural networks were trained to recognize patterns of amino acid concentrations as biotic or abiotic. We employed two-layer networks with twelve input nodes (one for each amino acid in the data set), three or four hidden-layer nodes, and two output nodes. Given an input pattern, each output node of the trained network generated a value between 0 and 1, which represented the Bayesian probability that the input pattern falls in the corresponding class. One output node was intended to register a “1” for biotic specimens, and the other was intended to register a “1” for abiotic specimens. In addition to using the twelve amino acids as inputs, we trained separate neural networks to classify the data as represented only by the ratios of elements H/C, N/C and O/C; in these cases the networks had only three input nodes, but were otherwise identical. With this approach, there are legitimate concerns about a specific network's ability to correctly extract the relationships within the data. If the data are sparse and/or the network is too large, it is possible that the network can “overtrain”; that is, memorize the specific instances of the training data rather than learn the underlying relationship between the categories. This is equivalent to generating a mathematical model with more adjustable parameters than the number of data points available: the experimenter cannot be certain that the model has extracted useful information about the system. In general, the number of data points should be much greater than the number of connections in the network. Several techniques are generally employed to prevent overtraining. Ideally, the experimenter simply gathers sufficient data to force the network to generalize. In practice, however, sources of data are often limited. In this study, only a very few examples of non-biologically-synthesized amino acids were available in the literature. In lieu of more data, we made numerous copies of the available data and added normally distributed noise to the copies. This approach is relatively insensitive to the actual amount or scale of the noise that is added. We also employed data set splitting, in which the training examples are split into two groups. The first group is used to train the network normally, while the second group is used only to observe the performance of the network. In general, the total error over a data set will decrease as the network correctly learns the relationships within the data. However, if the network begins to overtrain and learn the specific instances of the data in the first set, the error observed in classifying the second set will begin to increase. By watching for such behavior, it is possible to identify cases where the network fails to generalize and redesign the system to avoid the problem in the future.

5.5 Results

5.5.1 PCA results

Our amino acid data set was composed of twelve measurements (amino acid concentrations) for each data point. In raw form, these data are difficult for an experimenter to analyze visually (envision Fig. 5.1 with dozens of curves). When reduced to two or three principal components, however, visualization of the entire data set becomes tractable, as shown in Figs. 5.3.

Figure 5.3 shows a clear distinction between the biotic and abiotic categories. This result is seen, despite the fact that the PCA process does not seek to differentiate categories of data — only to isolate and represent sources of overall variance. The Synthesis and Murchison data tend to group together, indicating an affinity and supporting the hypothesis that the Murchison amino acids are abiotic. That the Murchison data group with each other and near most of the other abiotic data supports the hypothesis that any biological contaminants present are at a sufficiently low level that they do not mask the basic abiotic nature of the meteorite. Also, the protein data show greater variability than the sediment data, but in general

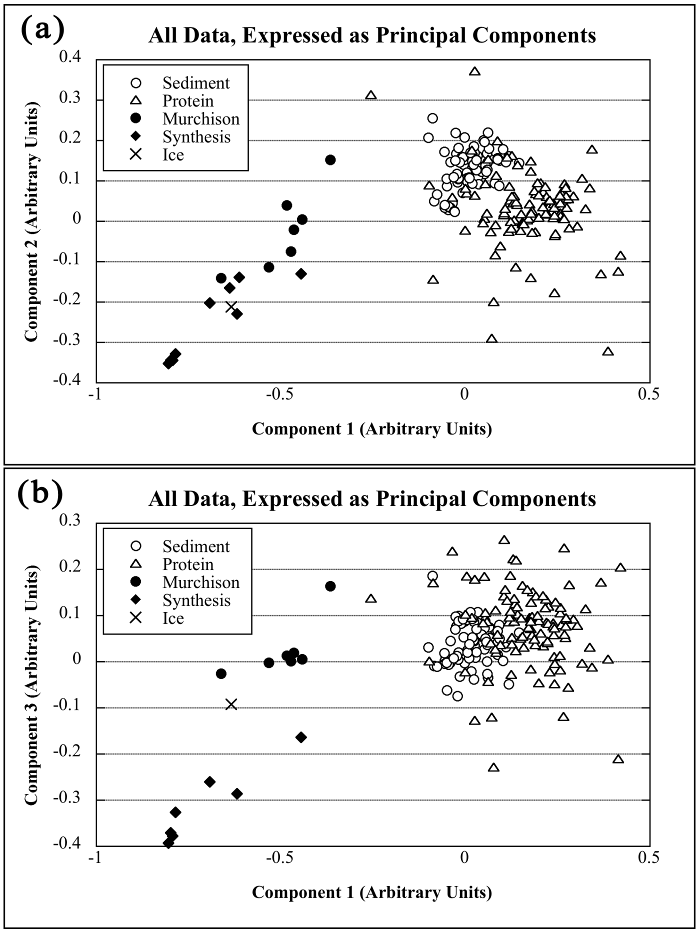


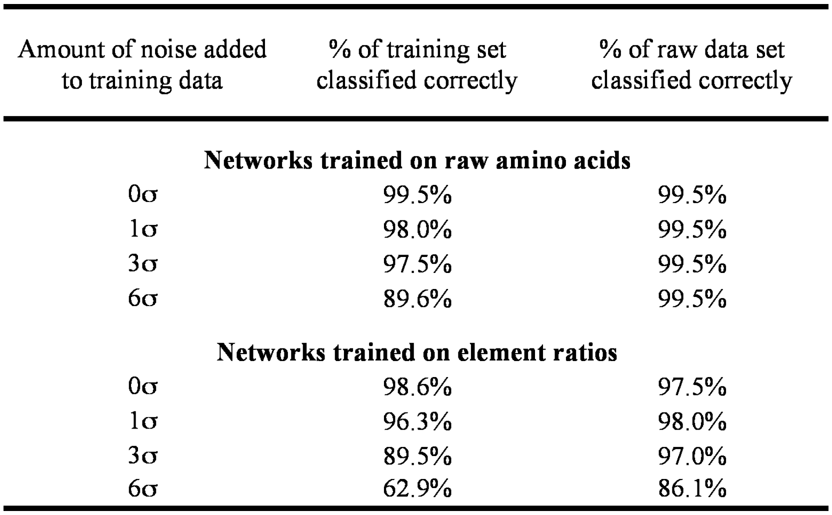
Figure 5.3: The data set plotted against various principal components. a. Component 2 vs. 1, representing 38% of the total variance present in the original data. b. Component 3 vs. 1, representing 34% of the total variance.

Component	**1**	**2**	**3**	**4**	**5**	**6**
Contribution to total variance	22.9%	14.9%	11.1%	9.4%	8.4%	7.7%
ALA	-0.486	*0.037*	**0.705**	*-0.062*	*-0.099*	0.166
ARG	**0.509**	-0.182	-0.265	-0.372	-0.189	*-0.034*
ASP	*0.061*	**0.739**	*-0.106*	*0.067*	0.459	*-0.141*
GLU	-0.234	**0.686**	0.264	-0.384	*0.110*	*-0.055*
GLY	**-0.808**	-0.356	-0.404	-0.164	*-0.085*	*0.029*
ILE	0.472	-0.322	-0.225	-0.285	**0.588**	*0.122*
LEU	**0.558**	-0.177	0.267	*-0.112*	*-0.011*	**0.635**
LYS	**0.503**	-0.385	0.299	-0.307	*-0.012*	-0.468
PHE	0.389	0.445	-0.200	-0.334	**-0.528**	*-0.061*
SER	**0.554**	0.253	*-0.155*	0.486	*-0.139*	*-0.016*
THR	**0.572**	0.320	*-0.005*	0.349	*-0.043*	0.208
VAL	0.466	-0.244	0.424	0.452	*0.040*	*-0.314*

Table 5.1: Composition of principal components and contribution of each to the variance of the original data set.

the sediment data are located near the center of the protein data. This suggests that amino acids extracted from life-bearing environmental samples and most proteins may group together. We expect some shift in composition due to biases in the extraction process; types of biological matter associated with different sediments and particle sizes; and breakdown of amino acids in the environment, but this effect appears to be small in these data. Analysis of the contributions of each amino acid to the principal components lends valuable information about the importance of each in the data set. Table 5.1 shows the contributions of each principal component to the total data set. The first three components together encode 48.9% of the variance of the original twelve-variable data set.

Within these components, glycine, aspartate, and glutamate show the strongest representation, indicating that the largest fraction of the statistically useful information in this particular dataset is encoded in those amino acids. This sort of analysis can assist in experiment selection and mission design.



Amount of noise added to training data	% of training set classified correctly	% of raw data set classified correctly
Networks trained on raw amino acids		
0σ	99.5%	99.5%
1σ	98.0%	99.5%
3σ	97.5%	99.5%
6σ	89.6%	99.5%
Networks trained on element ratios		
0σ	98.6%	97.5%
1σ	96.3%	98.0%
3σ	89.5%	97.0%
6σ	62.9%	86.1%

Table 5.2: Performance of ANNs at classification task.

5.5.2 ANN results

Regardless of the amount of noise added to the training dataset, ANNs trained on both the raw amino acid data and the data represented as elemental ratios were able to classify the data set with high accuracy. Networks were trained many times on each set of data to characterize the training process. In each case, after characterizing the training process for each set of training data, we trained a single representative network and measured its performance.

The lower accuracies observed in the "training data" column of Table 5.2 is an expected result, particularly as more random noise is added to the dataset. In some cases, the categories begin to overlap significantly and not all points can be correctly classified. However, the network makes a compromise; the result is an accurate generalization of the distribution underlying the actual data. The "perfect" results observed when the network is trained on noiseless data represent some degree of overtraining, made possible by the sparseness of the abiotic data.

Figure 5.4 shows how a typical network classifies the entire input space. After training on the input data, the network was used to classify a hypothetical set of data that spanned the entire glycine vs. alanine space. The output of this classification task was used to shade the background of a region on which the data set is superimposed. Black areas are categorized as biotic and white areas as abiotic.

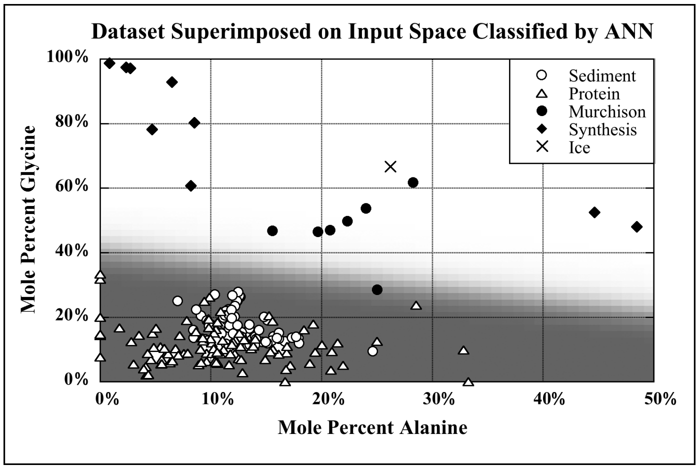


Figure 5.4: The dataset plotted as glycine vs. alanine, superimposed on a field representing ANN classification of the entire input space. Black regions represent areas classified as biotic, white regions represent abiotic areas. Intermediate levels represent partial confidence in classification.

This demonstrates how the network extrapolates beyond known data; a hypothetical sample that contained 50% alanine and 5% glycine (with the remainder more or less evenly distributed between the other 10 amino acids) would be classified as biotic. The decision boundary is actually sharper than it appears in Fig. 5.4. The network was trained on all twelve amino acids, but this figure demonstrates the classification of hypothetical data which did not include useful information about the other amino acids (the other ten input variables were set to identical average values for all points). A multi-dimensional field including axes for the other amino acids would show a much clearer boundary.

Like PCA, ANNs can provide information about the utility of each input measurement. We performed a first-order analysis of the importance of each amino acid by systematically training neural networks on copies of the data set with one variable removed. The changed ability of the resulting network to classify the remaining data indicates the value of the missing variable. Figure 5.5a shows the degraded performance when information about the relative abundance of glycine or alanine are withheld. Note that without information about the relative concentration of glycine, the networks have difficulty classifying the data set with better than 15% confidence; removal of alanine shows a similar effect. Removal of any other single variable has little impact on the network's performance. We also measured the effectiveness of the ANN algorithm when information on the relative concentration of only one amino acid (i.e., glycine as a molar percentage of the total amino acids present, but no other information) is available. Figure 5.5b shows the results. Information about the relative concentration of glycine is sufficient to make reasonable guesses about the nature of that sample. Some other amino acids provide a small amount of information, while others provide very little or no useful information. Aspartate, for example, provides just enough information to classify 40% of the set, but with very low confidence. The relative concentration of glutamate, on the other hand, indicates nothing about the biotic or abiotic nature of the sample. Alanine concentration is a weak classification criterion by itself.

Non-biological syntheses of amino acids generate large (50% to 80%) quantities of glycine [61,59]. This makes intuitive sense, as glycine has a low formation energy. In biotic samples, however, glycine will generally constitute a much smaller fraction of the total, as living systems cannot manufacture useful proteins that are 80% glycine.

5.6 Conclusions

Techniques like PCA and ANN show potential for use in automated scientific analysis by remote probes. In order to improve their usefulness for autonomous and non-Earth-centric life detection, it will be necessary to expand and generalize upon the techniques used in this study. In particular, the employment of "unsupervised" pattern recognition systems which classify and categorize incoming data without reference to a training set of known examples is indicated. We envision an automated science platform making mission decisions by classifying samples that have undergone simple analysis. With an appropriate classification, an optimal subset of samples can be selected for more detailed analysis by other, presumably more time-consuming, experimental tools. The ANN technique can enable powerful and robust analysis of this sort of data. While many of the individual amino acids, such as aspartate, allow for only a very weak classification, the network can combine multiple weak

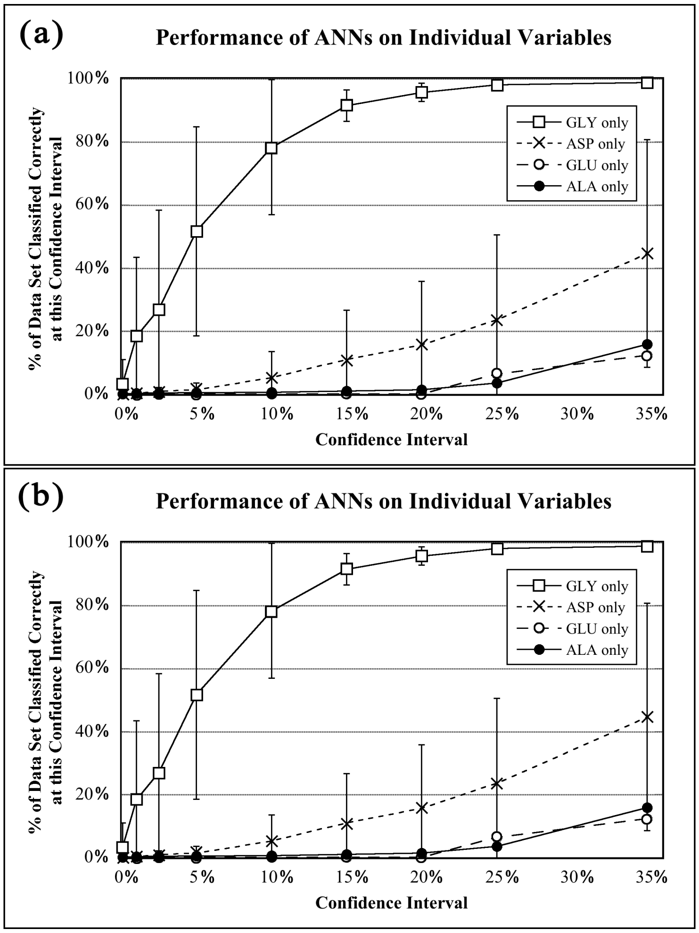


Figure 5.5: Confidence curves demonstrating the changes in classification performance as data are removed from the training set. a. individual input variables are removed; the result is compared to the result with the complete dataset. b. only single input variables are provided to the network during training.

measurements to produce a high-confidence classification in ways that would not be obvious to a human observer. Both ANNs and PCA share a strong robustness to noise — an important feature — as autonomous systems will be subject to many sources of signal degradation. These systems can provide some level of buffer, even against terrestrial biological contaminants as long as the contaminant level is not high enough to grossly distort the shape of the curve being measured. It is important to note that the ANNs classify the H/C:N/C:O/C representation of the data set nearly as well as the raw amino acid concentrations. This result implies that, for Earth-like life at least, the elemental ratios may be sufficient for a high-confidence discrimination. This is primarily because abiotic syntheses produce such a large abundance of glycine, which has very high N/C and O/C ratios relative to the other amino acids. A probe may therefore be able to efficiently evaluate the amino acids present in a sample with a simple mass spectroscopy experiment rather than the liquid chromatography and other equipment necessary for quantification of individual amino acids. The argument that high percentages of glycine are not adequate for stable polypeptide structures may even imply that this analysis is sufficient for detection of any living system that uses polypeptides, even if it is otherwise unlike terrestrial biochemistry. We believe that the kinetic and thermodynamic constraints of biochemistry will also produce similar biosignatures in other chemical species, some of which may be more easily isolated and measured. The element ratios of the bulk hydrolysable fraction or bulk volatile fraction of a sample may contain sufficient information for these techniques to make the biotic/abiotic discrimination. In this study, we were constrained by the availability of amino acid data in the literature, in which the protein amino acids represent the lowest common denominator of amino acids examined in all the studies. It would be instructive to include amino acids such as sarcosine that are sometimes found in high quantities in non-biological syntheses [61], but are not expected to appear in abundance in terrestrial environments; and methionine and cysteine, which are universal in known living systems, but absent in abiotic syntheses. In future studies, it would be interesting to compare patterns from biological sources that have undergone extensive environmental degradation while remaining protected from contamination by recent sources, such as the interior of fossils or amber-encased insects as described in [8]. We would also like to compare these measurements to extracts from unfractionated, intact biological material. We envision empowering automated systems with the ability to make simple predictions about the expected abundances of potential biomarker chemicals based on basic knowledge of an environment and thermodynamic principles. For example, in most environments, if non-biological synthesis is the dominant process for amino acid abundance, we would expect to find relatively large quantities of the easily formed amino acids glycine, alanine and perhaps sarcosine, and little else. Comparison with these predictions would give an *in situ* platform a basis for tagging "unusual" samples for further analysis. This capability is essential if the system is to be capable of detecting life forms even if their biochemistry differs from that of terrestrial biota.

5.7 Acknowledgements

The research described in this paper was carried out at Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The authors thank P. Conrad for useful discussion.

Chapter 6

Pattern Recognition for Non-Earth-Centric Detection of Chemical Biosignatures

To be submitted to *Astrobiology*, February 2005.

Authors as published: Evan D. Dorn, Christoph Adami

6.1 Abstract

We demonstrate a method for automated detection of chemical biosignatures that can, in principle, detect extraterrestrial life forms even when their biochemistry is both unknown and dissimilar to terrestrial biochemistry. Because organisms construct cellular machinery to synthesize the compounds they need for survival, we expect compounds used actively in biotic metabolisms to appear at concentrations that stand out from the thermodynamics- and kinetics-dominated background. By examining the relative concentrations of multiple chemicals, we can deduce whether their observed concentration ratios fall outside experimentally determined reasonable ranges for abiotic synthesis. A trained radial-basis function neural network can compute the extent to which an individual measurement of multiple concentrations differs from known abiotic patterns, irrespective of its similarity to measurements from the terrestrial biosphere. We demonstrate the technique as applied to a published body of measurements of amino acids from a variety of sources.

6.2 Introduction

Astrobiology suffers from the so-called “one data point” problem, in that we have only one example of an evolved biochemistry and cannot easily perform experiments to determine the range of potential biochemistries. Therefore, a good life-detection experiment should satisfy two criteria:

1. Test for a phenomenon that is an inevitable result of processes universal to life, regardless of substrate or biochemical particulars.

2. Detect the signs of that phenomenon, regardless of its specific similarity to the terrestrial biosphere.

We demonstrated previously [33] how a neural network could discriminate distributions of amino acids from life-bearing (biotic) and non-life-bearing (abiotic) samples. The ability of an unsupervised computer to make discriminations of this sort will become important as robotic missions with astrobiology components begin to range further and increase in number and therefore require greater autonomy. However, the previous system fails criterion (2) above because it diagnoses amino acid biosignatures based on their similarity to known measurements from terrestrial sediments. Here, we demonstrate an improvement on the previous technique that reduces the dependence on knowledge of existing biochemistry.

6.3 Life's effect on environmental chemistry

We have examined elsewhere [34] how organsims leave a chemical biosignature in their environment as natural selection induces them to catabolize particular compounds at rates that meet fitness criteria. Here, we seek to formalize the concept.

The equilibrium concentration, c_i, of particular chemical i in an environmental sample is described by some complex function:

$$c_i = F_i(k_{f,abiotic}, k_{f,biotic}, k_{b,abiotic}, k_{b,biotic}, \dots) \tag{6.1}$$

where $k_{f,abiotic}$ is the rate of abiotic synthesis or inflow, $k_{f,biotic}$ is the rate of biotic synthesis, $k_{b,abiotic}$ is the rate of abiotic breakdown or diagenesis, and $k_{b,biotic}$ is the rate of biologically catalyzed breakdown and consumption. If i represents a bioactive molecule, $k_{f,biotic}$ and/or $k_{b,biotic}$ are nonzero in the presence of metabolizing organisms and thus, we may expect c_i to take on a different range of values in the presence of life as compared to sterile environments. In practice, however, the function F and these terms are unknowable. In addition, variables may affect the overall scale of measurements c_i : dilution, temperature, availability of reaction precursors, etc. The concentration of a single molecule in the absence other information is not generally very useful.

However, useful information may be encoded in the relative concentrations of several species $c_1, \dots, c_M$, particularly of a family of related chemicals. The ratios $c_1 : c_2$, $c_1 : c_3$, etc. are invariant with respect to overall dilution and may tell us about these species' formation dynamics.

When life is not present, the ratio of two concentrations reflects the degree to which thermodynamic and kinetic processes favor the formation and persistence of one species over another. For example, it is generally assumed that in the absence of life the concentration of O_2 in most atmospheres will be low, relative to CO_2 and N_2 due to oxygen's high reactivity and tendency to bind to minerals. The example is trivial, but serves to illustrate the universality of the statement. We are most interested in small organic molecules such as amino acids, fatty acids, polycyclic aromatic hydrocarbons, etc., because they form in both biotic and abiotic conditions and are generally implicated in the origins of life. Experiments in abiotic synthesis and measurements from abiotic sources such as carbonaceous chondrites invariably show a distribution of organic molecules that strongly favors small compounds. This makes sense: larger molecules generally have higher formation energies, have a greater number of possible conformers, and require more steps in their synthesis pathways, so the formation rate of larger species should, on average, be smaller.

Life, on the other hand, drastically alters local chemistry in the struggle for survival. By concentrating reactants inside a membrane, constructing highly specific catalysts (enzymes) to reduce activation barriers, and expending energy to drive otherwise unfavorable reactions, biota synthesize a specific set of molecules at extremely high rates. Given a set of concentrations $c_1 \ldots c_M$, it is likely that some ratios within that set will be anomalous with respect to abiotic synthesis if any of the c_i's are involved in biotic metabolism and life is present.

Since biosynthesis and the genetic code that governs it are the product of natural selection, we can describe this phenomenon as selection acting indirectly on the chemical composition of an environment. Just as selection acts to encode information about a fitness landscape in a genome by perpetuating genes that improve adaptation, so does it encode similar information in the chemical composition of an environment by perpetuating genomes that synthesize molecules at the most adaptive rates.

Therefore, chemical concentration ratios that are anomalous with respect to abiotic formation dynamics may indicate that another process — selection — is acting on bulk molecular synthesis in the environment. Since selection is expected to be a feature of all living systems, this phenomenon may serve as a universal life diagnostic. We most often study this process in small molecules that frequently polymerize or otherwise combine to build biomass, so we refer to them generically as “monomers” and the effect life has on their abundance ratios as the “Monomer Abundance Distribution Biosignature” (MADB).

We approach detecting an MADB by considering a series of concentration measurements $c_1 \ldots c_M$ from a single sample, where $1 \ldots M$ represent members of a related chemical family. Described as a vector $\boldsymbol{c}$, the ratios of this set of measurements can be examined by considering either the unit vector $\hat{\boldsymbol{c}}$, or the vector scaled relative to one of its components c_m:

$$\boldsymbol{c}_{\text{scaled}} = \frac{\boldsymbol{c}}{c_m} \tag{6.2}$$

Whether we choose $\hat{\boldsymbol{c}}$ or $\boldsymbol{c}_{\text{scaled}}$ depends on the particular family of measurements in question and whether or not an appropriate c_m is available, and it does not matter to the design of the algorithm. Henceforth, we will use $\boldsymbol{c}$ to mean either of these two formulations.

Detecting the MADB then becomes a problem of determining whether a new measurement vector $\boldsymbol{c}$ belongs to the population of abiotically generated monomer abundance patterns $\mathbb{C}_{\text{abiotic}}$ or the population of biotic patterns $\mathbb{C}_{\text{biotic}}$. Previously [34], we demonstrated how a neural network can be trained to make this discrimination. Here, we improve on the system and significantly reduce its dependance on knowledge of known biochemistry by searching for patterns $\boldsymbol{c}$ that are not members of $\mathbb{C}_{\text{abiotic}}$, rather than by searching for patterns which are members of $\mathbb{C}_{\text{biotic}}$ (as presently understood).

While it is not technically necessary for all the monomers to be from the same family, in practice, related chemicals often share similar formation mechanisms and so the distribution of possible vectors within $\mathbb{C}_{\text{abiotic}}$ is more constrained and can be better characterized. This is important, since our ability to detect the MADB depends crucially on our ability to fully characterize the range of $\mathbb{C}_{\mathrm{abiotic}}$. In this study, we use amino acids as our test case.

6.4 Amino acids

Amino acids are a well-studied chemical family fitting the criteria for the MADB. We use a published body of data about amino acids from both biotic and abiotic samples to design and test our system. They have been studied before as a biosignature for astrobiology [7, 6], and can be synthesized by both biotic and abiotic processes. Mixtures synthesized by nonbiological processes tend to be almost entirely low-molecular-weight and low-energy-of-formation species such as glycine, alanine, beta-alanine, gamma-aminobutyric acid, and sarcosine. Samples from terrestrial sediments, however, generally exhibit moderate levels of at least all twenty protein amino acids including high-molecular-weight and high-energy-of-formation residues that are never seen in abiotic synthesis. This reflects the basic process described above: histidine (for example) is needed for survival and competition, and therefore, organisms will manufacture enzymes and expend energy to increase $k_{f,biotic}^{HIS}$ to significant levels.

Fifty years of experimentation attempting to model the formation of amino acids on the prebiotic Earth or other bodies, and measurements of the amino acid contents of carbonaceous chondrites, have endowed us with a wealth of data on amino acid expression in abiotic synthesis. Several recent publications in the field of organic matter cycling provide amino acid quantification of samples from a known biosphere. In combination, these provide a basis data set for constructing pattern-recognition networks in this work. A complete list of sources may be found in Table 6.4.

These disparate publications often differ in the chromatography techniques used, the amino acids reported, and the units used. We used the following procedure to combine them into a consistent database:

1. Data were extracted from figures, where necessary, by scanning the publication at 1200 DPI and measuring the graphics in Adobe Photoshop.

2. Data were converted to mole units if necessary.
3. Each measurement was entered into a row of a database that had columns for 31 different amino acid residues. Columns for unreported residues were left blank.
4. Residues reported as coeluted were both given the reported value.
5. Where D- and L- enantiomers (mirror-image stereoisomers) were reported separately, their sum was used.
6. Values described as "approximate" or "estimated" were used without change.
7. The upper bound was used for small values (e.g., ALA < 0.1).
8. Values reported as "trace" were assigned a zero.
9. Residues described in the text as "not found" or "not seen" were assigned a zero.
10. If the text contained a clear indication that the listed values were exhaustive (e.g., "the table lists the concentration of all observed amino acids" or "other amino acids were present in trace quantities"), a zero was entered into each remaining column.

A subset of twelve columns was then selected that contained the greatest fraction of residues to which a definite value could be assigned: GLY, ALA, BALA (beta-alanine), ASP, GLU, SER, VAL, LEU, ILE, LYS, PHE, and THR. Finally, each vector of twelve amino acid concentrations was divided by its normal to generate a unit vector.

6.5 Non-Earth-centric life detection

Since our goal is to detect life even where it may differ from terrestrial biochemistry, we want to make no assumptions about the population of possible biotic patterns $\mathbb{C}_{\text{biotic}}$. The system we described in [33] is limited because it implicitly makes such

#	Energy Source	Atmosphere or Substrate	Source
Abiotic Syntheses			
1	Arc-Discharge	H_2, CH_4, NH_3, H_2O	[62]
1	Arc-Discharge	H_2, CH_4, NH_3, H_2O	[63]
1	Arc-Discharge	NH4+, NH_3, N_2, CH_4, H_2O	[74, 95]
6	Arc-Discharge	H_2, N_2, H_2O plus (CO),(CO_2), (CH_4), or (CH_4, NH_4^+)	[76]
1	DC-Discharge	N_2, CH_4 (9:1) "Titan Tholin"	[47]
5	DC-Discharge	CH_4, NH_3, H_2O "Jupiter Tholin"	[57]
1	DC-Discharge	N_2, CH_4 (999:1) "Triton Tholin"	[59]
Abiotic Meteorites			
1	Murchison		[50]
1	Nagoya		[22]
4	Yamato-74662	Interior/Exterior, Hydrolyzed/Unhydrolyzed	[83]
1	Murchison		[27]
2	ALHA 77306	Interior/Exterior Hydrolyzed/Unhydrolyzed	[27]
2	Murchison		[37]
1	Murchison		[36]
1	Murchison		[35]
Terrestrial Sediments			
13	Sediment Trap, Saanich Inlet, B.C. 13 depths, 0-80 cm		[18]
1	Porcupine Abyssal Plain sediment core, 1-2 cm		[44]
19	North Sea sediment cores and traps		[28]
9	Laurentian Trough suspended particles and sediments		[17]
46	Washington coast, suspended particles from five sites w/6-9 size fractions ea.		[46]
24	Amazon River particles, 8 locations w/3 size fractions ea.		[42]
9	Washington coast, 4 locations and 5 organic matter types		[45]

Table 6.1: Sources of data used to train the neural network.

an assumption. It uses an artificial neural network (ANN) with sigmoidal neurons to distinguish two clusters of data from different populations residing in a high-dimensional space. Presented with a novel case, the ANN described previously will report which of the two clusters matches more closely.

Such a system raises a concern for non-Earth-centric life detection. If a hypothetical extraterrestrial organism happened to use a significantly different set of amino acids from those used on Earth, the system would be unprepared to handle a sample from that environment. In other words, because the system was trained to classify a novel input based on its similarity to two known populations, its response to a sample from a third, previously unknown population is undefined. Here we improve the technique by employing a different sort of classifier that discriminates based upon the overall distance (in a difficult to reduce nonlinear sense) from a multidimensional cluster. Novel samples will be classified as "life" if their measurements look sufficiently different from the characterized population of nonliving samples. A cartoon comparison of the previous technique and the present one can be seen in Figure 6.1.

We define a system that uses a known set of N abiotic patterns $p_1 \dots p_N$ (represented henceforth by the N-by-M matrix $\boldsymbol{P}$), as an estimate of the range of $\mathbb{C}_{\text{abiotic}}$ and flags as potentially biotic those incoming samples that fall outside that range. For a single measurement c_m in one dimension, this is equivalent to a statistical t-test computing the probability that c_m falls outside the population sampled $\boldsymbol{p}$. Neural networks give us an approach to computing a similar test where our measurements are multidimensional and the population $\mathbb{C}_{\text{abiotic}}$ may have a complex shape in N-space.

Such a system should in principle be capable of detecting life even where it differs significantly from terrestrial biota. Designing a general system for detecting life thus becomes a problem of choosing a set of measurements $C_1 \dots C_n$, at least some of which are likely to be affected by the presence of any kind of life, and thoroughly characterizing the distribution of $\mathbb{C}_{\text{abiotic}}$.

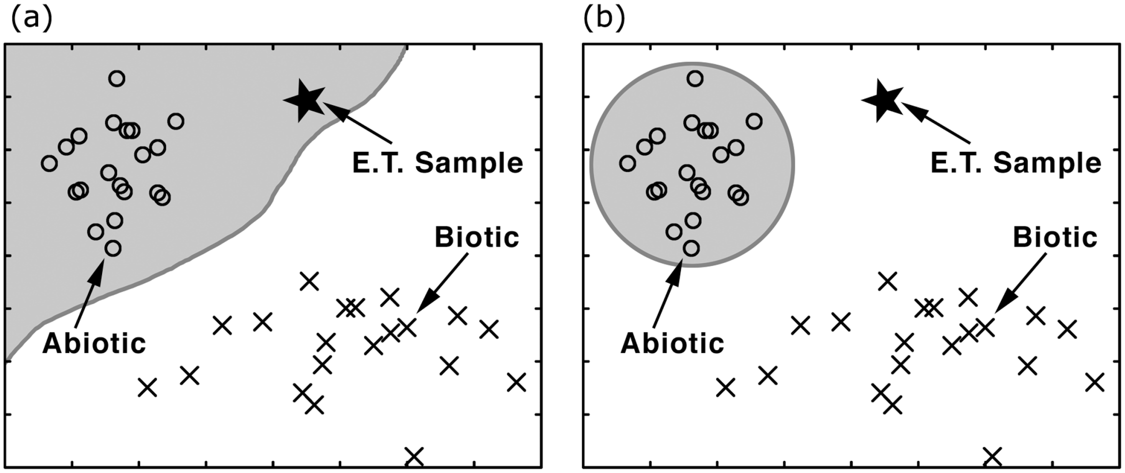


Figure 6.1: Responses of two hypothetical neural networks. Samples falling within the shaded area will be classified as abiotic. In (a), a traditional ANN splits the input space into two regions based on the training inputs. A novel sample from a different population might be classified incorrectly. In (b), a radial-basis-function neural network has been trained to characterize the abiotic class based on known data. Points falling outside the abiotic scope are classified as "biotic" even if they don't match known biotic data.

6.6 Methods

We use a variation of classic radial-basis-function (RBF) neural networks (see [10] for an introduction). Such a network employs a layer of O "radial-basis neurons": computational units, each of which effectively defines a point in N-space and responds to input data near that point. Neuron i produces a response g_i to an input pattern in the column vector $\boldsymbol{c}$:

$$g_i(\boldsymbol{c}) = r(\|\boldsymbol{w}^T{}_{1,i} - \boldsymbol{c}\| b_{1,i}) \tag{6.3}$$

.

Equivalently:

$$g(\boldsymbol{c}) = r(\boldsymbol{w}_{1,i}\boldsymbol{c}b_{1,i}) \tag{6.4}$$

where $\boldsymbol{w}_{1,i}$ is a row vector representing the ith unit's center in N-space and the bias

$b_{1,i}$ (or "spread") is a constant that sets its radius. The transfer function r is defined as

$$r(x) = e^{-x^2}. \tag{6.5}$$

With a layer of O neurons, we represent their combined $\boldsymbol{w}_{1,i}$'s as the N-by-O matrix $\boldsymbol{W}_1$, their biases as the vector $\boldsymbol{b}_1$, and their outputs as the vector $\boldsymbol{g}$. The values of $\boldsymbol{W}_1$ are called the weights due to their similarity to the weights of a traditional ANN.

The second layer of the network consists of a layer of linear nodes that each compute a weighted and biased sum of the outputs of the previous layer, the weights being specified by a second matrix $\boldsymbol{W}_2$. The response of the entire network to an input vector $\boldsymbol{c}$ is then:

$$O(\boldsymbol{c}) = \boldsymbol{W}_2\, r(\|\boldsymbol{W}_1\boldsymbol{c}\|\boldsymbol{b}_1) + \boldsymbol{b}_2. \tag{6.6}$$

Given a sufficient number of nodes in the first layer, such a network may be used to approximate any function. Choosing the values of $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ is called "training" the network, and is usually done using an iterative gradient descent function. In addition, one other parameter must be chosen: M, the number of neurons in the first layer.

Here, we define a function of the chemical measurement vector $\boldsymbol{c}$:

$$\mathcal{F}(\boldsymbol{c}) = \begin{cases} 1 & \text{if } \boldsymbol{c} \text{ is in } \mathbb{C}_{\text{abiotic}} \text{ and} \\ 0 & \text{otherwise,} \end{cases} \tag{6.7}$$

and train a radial-basis network with a single output-layer node to estimate this function using known abiotic example vectors $\boldsymbol{P}$. This function can be fit by an RBF network without any examples in the biotic, ($\mathcal{F} = 0$) region, by requiring the bias vector $\boldsymbol{b}_2$ to be zero. Since the output of the neuronal response function in Eq. 6.3 is significant only when its input is near the vectors described in $\boldsymbol{W}_1$, the effect of setting $\boldsymbol{b}_2 = 0$ is to force the network to output zero everywhere except near its

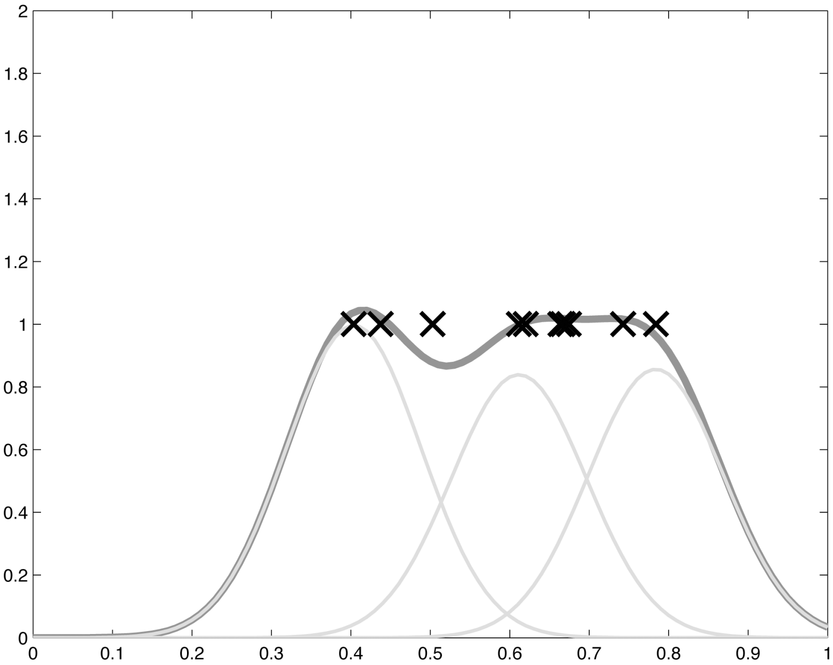


Figure 6.2: Response of an RBF network trained on a few sample data points generated randomly in the range 0.4 to 0.8. The network uses the sum of three zero-based gaussian kernels (shown in light gray) to closely fit a plateau function matching the data.

neurons' response centers, which we will train to approximate the the range of the known biotic points in $\boldsymbol{P}$. Figure 6.2 shows how the system fits a simple function with one input where the target class extends over the range $0.4 - 0.8$.

Setting the number of neurons M and the width of the responses (defined by $\boldsymbol{b}_1$) correctly is important. If M is too large or the neurons are too narrow, the network can "overtrain," that is, learn the specific individual points in the data set rather than the general distribution. In general, M should be less than the number of data points in the training set and the width of the neurons s (where $s = 0.8325/b_1$; 0.8325 is the value at which the transfer function $r(x)$ has dropped to 50% of its maximum, $r(0.8325) = 0.5$) should be large enough to prevent the system from fitting nodes to single data points. If, on the other hand, the spread s is too large, the response of the network will not drop off quickly outside the range of the target class and points outside $\mathbb{C}_{\text{abiotic}}$ will generate an incorrect output.

In order to maximize the non-Earth-centric nature of this pattern-matching system, we want to train the network in a way that is maximally agnostic of the values of $\mathbb{C}_{\text{biotic}}$. Given a set of training patterns $\boldsymbol{P}_{\text{abiotic}}$ and $\boldsymbol{P}_{\text{biotic}}$, we use the following procedure: any particular value of O, the centers ($\boldsymbol{W}_1$) and magnitudes ($\boldsymbol{W}_2$) of the neurons' responses are computed using a gradient descent function that seeks to minimize the error over the set of abiotic training input patterns $\boldsymbol{P}_{\text{abiotic}}$ with no knowledge of the biotic data set. We iterate this procedure over a range of s and keep the network that minimizes the mean squared error over the combined set of inputs $\boldsymbol{P}_{\text{abiotic}}$ and $\boldsymbol{P}_{\text{biotic}}$. In effect, we are training the network to characterize $\boldsymbol{P}_{\text{abiotic}}$ and only using $\boldsymbol{P}_{\text{biotic}}$ to set the width of the response $\boldsymbol{b}_1$, our threshhold for "how far away" a measurement must be from the abiotic class to be declared "biotic". Figure 6.6 shows the generalized response patterns of several trained networks.

6.7 Results

Trained radial basis function neural networks with 8 to 15 neurons recognized the known pattern of amino acid abundance from abiotic sources and rejected terrestrial biotic samples with high reliability. Typically, all of the abiotic training data score at least 0.85, and all of the biotic training data score lower than 0.10, where a higher score represents the degree of similarity to known abiotic data. The degree to which the network prefers to accept false positives vs. false negatives is customizable by the researcher in decisions about training thresholds; we tended to err on the side of excluding unknown points from the abiotic set. Figure 6.7 shows the distribution of patterns in the training data, for comparison to the results.

We have used an implementation of this network — trained with twelve basis functions — to classify a a number of "unknown" samples that were not available in the training set to observe how well it generalizes. In general, these data were less well suited for the test than the training set: many are considered potentially contaminated, and most reported only a few of the twelve residues we included in the training database. Nonetheless, we find that the network can make a largely accurate discrimination between biotic and abiotic samples. Table 6.3 shows the outputs of

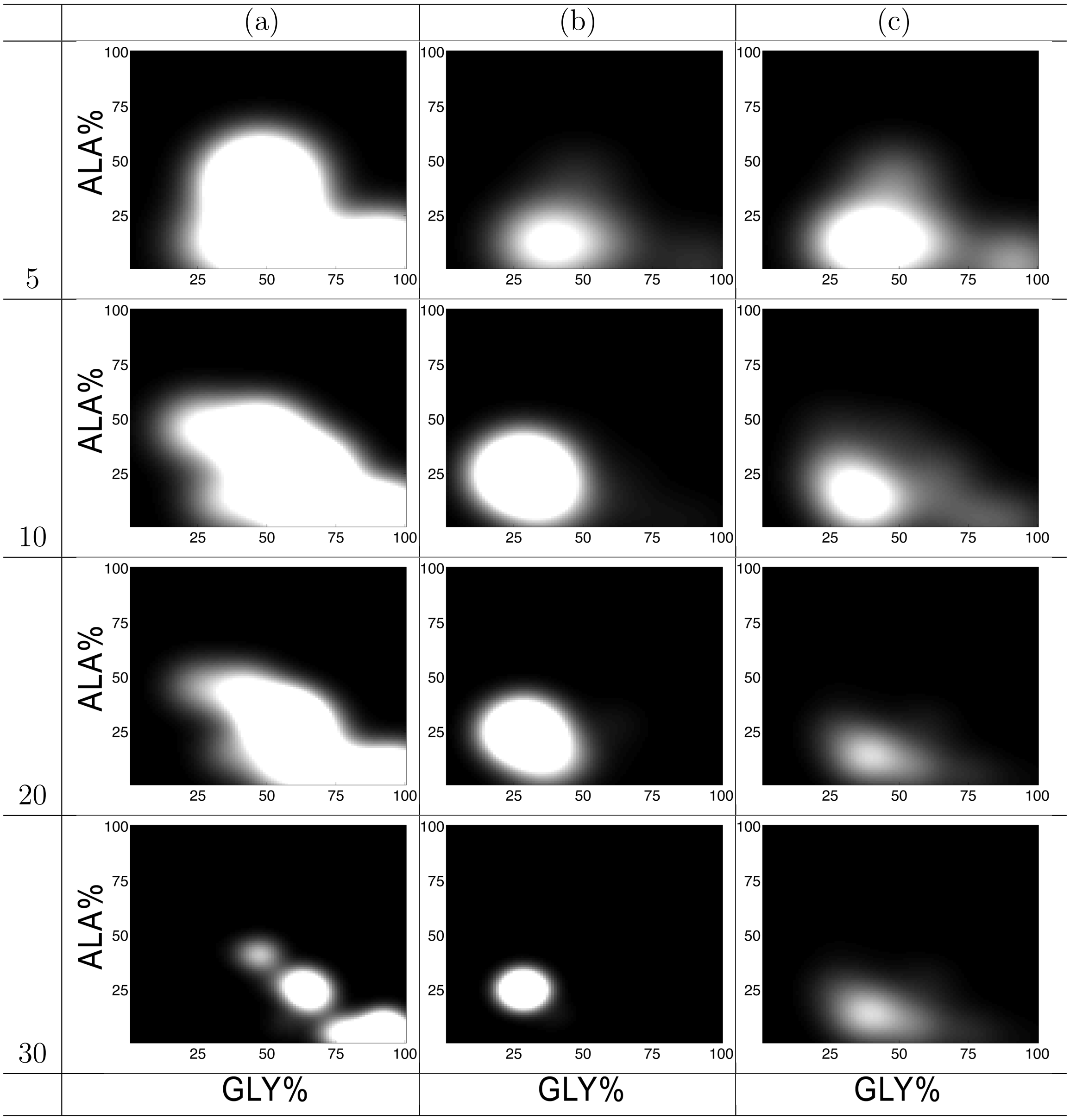


Figure 6.3: Response fields of several RBF networks, shown over the plane defined by ALA percentage of total amino acid content vs. GLY percentage. White areas match known abiotic data from the training set. Rows are networks trained with 5, 10, 20, and 30 radial basis functions, respectively. Columns show the response of the network over the ALA% vs. GLY% plane with (a) all other inputs held at zero, (b) other inputs held at their average over the abiotic training set and (c) other inputs held at their average over the biotic training set. Note how the system overtrains when 30 basis functions are used, learning the specific data points in the training sample rather than the overall response.

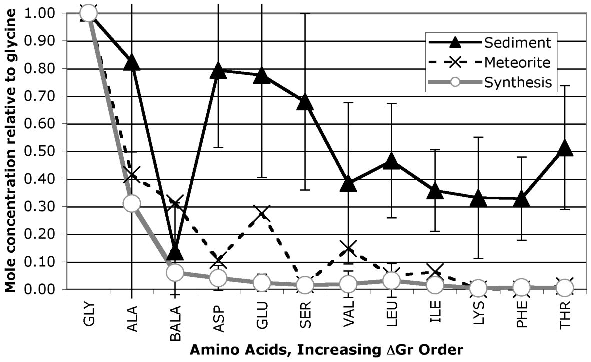


Figure 6.4: The abiotic data set used to train the neural network, represented as category averages, with known biotic measurements shown for comparison. Error bars are one standard deviation, to show the wide distributions encountered. Sediment n=121, Meteorite n=13, Synthesis n=18.

#	Description	Source
1	Abiotic synthesis via UV photolysis on ice in vacuum	[64]
1	Abiotic synthesis via proton irradiation in vacuum	[89]
5	CM chondrites: Murchison, Murray, Nogoya, Mighei, Essebi	[12]
2	CI chondrites: Orgueil and Ivuna	[12]
3	Other meteorites: Renazzo, Allende, Tagish Lake	[12]
2	Nakhla meteorite	[39]
4	Fossilized hadrosaur teeth	[72]
1	Body of amber-encased fly	[8]

Table 6.2: Sources of data used as unknowns to test the neural network.

the network when presented with twenty previously unseen samples for classification.

Munõz-Caro et al. [64] report the synthesis of amino acids via UV hydrolysis in a laboratory simulation of interstellar ice particles, and Takano et al. [89] report the synthesis of amino acids from an inorganic mixture via proton irradiation in a simulation of interstellar gas. Though the experimental conditions and temperatures are quite different from the spark-syntheses that generated most of the data in the training set, the network nonetheless groups both of these measurements with the other abiotic data with a high degree of confidence. Figure 6.7 shows these measurements compared to the training data.

Botta et al. report new measurements of amino acids from ten chondrites, most of which are believed to be at least partially contaminated by terrestrial amino acids,

Sample Description	Result
Murchison	**0.788**
Murray	0.674
Nogoya	0.358
Mighei	*0.251*
Essebi	*0.003*
Orgueil	*0.000*
Ivuna	*0.010*
Renazzo	0.346
Allende	*0.207*
Tagish Lake	*0.184*
Amber-encased fly body	0.340
Amber matrix	0.287
Nakhla (free fraction)	0.609
Nakhla (total amino acids)	0.580
Hadrosaur Teeth (1)	*0.079*
Hadrosaur Teeth (2)	*0.148*
Hadrosaur Teeth (3)	0.354
Hadrosaur Teeth (4)	*0.209*
UV-irradiated ice crystals in vacuum	**0.984**
Proton-irradiated gas mixture	**1.036**

Table 6.3: Output of the neural network when presented with novel patterns. The result represents the degree to which the network correlates this input pattern with the abiotic training set, which included samples from purportedly uncontaminated CM chondrites and classic spark-synthesis studies. The strongest matches (>0.7) are presented in boldface, and the weakest matches (<0.3) are presented in italics.

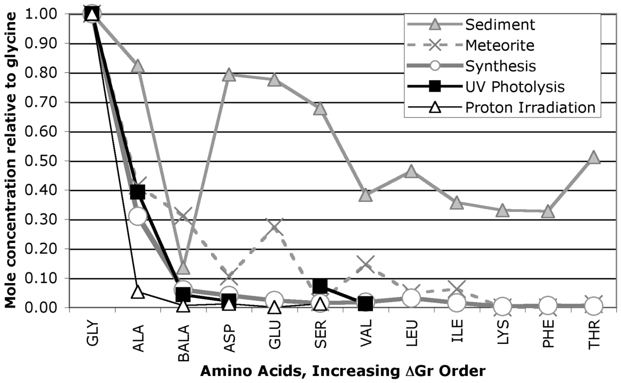


Figure 6.5: Data from two abiotic synthesis experiments not included in the training data set. "UV Photolysis" is a single profile from an experiment in which organic compounds were formed via UV irradiation on ice crystals in a vacuum chamber, while "Proton Irradition" is a measurement from an experiment simulating formation of organic compounds in interstellar gas. These profiles received the highest score, indicating a match to the abiotic training set of any in our study.

based largely on enantiomeric ratios [12]. They consider the presence of the amino acids isovaline and α-amino isobutyric acid (which are not found on Earth) to be diagnostic for extraterrestrial sources. These two amino acids were found at the largest abundances in the CM chondrite group, particularly in Murchison and Murray. Despite the fact that iVAL and AIB were entirely excluded from our analysis, this group produced the strongest abiotic signature of the meteorites, indicating that other features caused them to correlate with the abiotic training set (which included CM chondrites, including other samples of Murchison). Nogoya and Mighei both have lower scores than the other meteorites in this group, largely due to a high level of beta-alanine. Figure 6.7 compares this set of CM chondrite data to other data used in our study to illustrate the difference.

The network produced scores very near zero for the two CI chondrites, Orgueil and Ivuna, and for the CM chondrite Essebi. This is almost certainly due to those bodies' very high relative concentrations of beta-alanine (BALA/GLY > 2). It has been argued that the formation of large quantities of beta- amino acids seen in these three bodies results from a process that is abiotic, but distinct from the Strecker mechanism believed responsible for most amino acid formations in CM chondrites like Murchison and Murray [12, 11]. Though this process is abiotic, the behavior of our classifier is precisely what we would expect it to be, since no CI chondrites or other samples exhibiting that process were included in the training data. Thus, these measurements serve to demostrate how a classifier of this sort responds to truly novel data. Figure 6.7 compares the CI chondrites data to other data used in our study to illustrate the difference.

A measurement from the Renazzo CR meteorite [12] does not match the abiotic training set well, but has the highest value aside from the samples of Murchison and Murray, indicating some similarity in formation process. This is consistent with previous interpretations [12]. The low values for the samples of the Allende and Tagish Lake meteorites (from [12], shown in Figure 6.7) are consistent with other interpretations that amino acids in those bodies are likely due to terrestrial contamination [21, 12, 49].

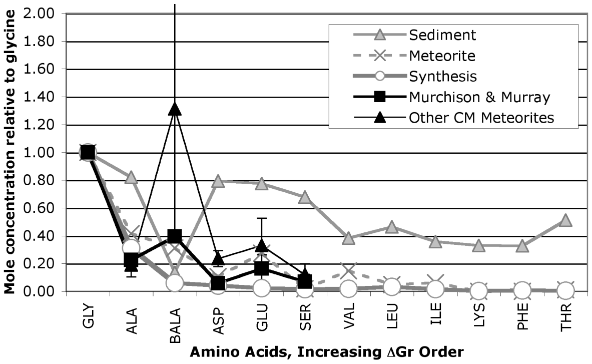


Figure 6.6: Data from CM meteorites not included in the training data set. The trace labelled “Murchison & Murray” is an average of the two measurements from those meteorites reported in [12]. The “Other CM Chondrites” is an average of the measurements from the Nogoya, Mighei, and Essebi meteorites reported in the same source. Error bars are standard deviation in each case.

Glavin et al. report a pair of amino acid series from the Martian meteorite Nakhla and determine that it is partially contaminated, based on enantiomeric excess [39]. Our results — which disregard optical activity — are consistent with that interpretation: the Nakhla profiles received intermediate scores from the neural network, representing the presence of a mixture of biotic and abiotic processes. While to the eye, the Nakhla measurements resemble the patterns seen in the CM chondrites used in the training set (Murchison, Yamato 74662, and ALHA 77306), the scores are lower than those for the new samples from Murchison and Murray described above. It is worth noting that the two extractions from Nakhla exhibit systematic bias inherent in the extraction method: relatively more alanine was extracted in the free (aqueous) fraction while realtively more glutamate was found in the total (acid-hydrolyzed) fraction. Nontheless, both patterns generate similar responses from the classifier, because amino acid patterns from both aqueous and acid-hydrolyzed fractions of meteorites were included in the training data (Table 6.4).

Bada et al. report measurements of six amino acids (ASP, SER, GLU, GLY, ALA, VAL) from a sample of 40 million year old amber and an insect body included in the amber [8]. They detect essentially no racemization or breakdown of the enclosed amino acids and conclude that the contents have been nearly 100% conserved. The results of the network are consistent with that interpretation, assigning fairly low

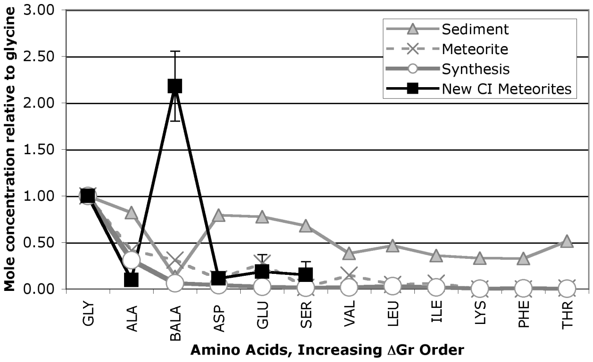


Figure 6.7: Data from CI meteorites not included in the training data set. Note the very high levels of beta-alanine relative to all other residues, including glycine.

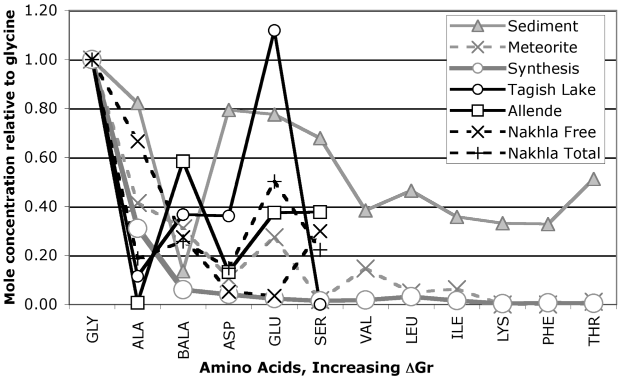


Figure 6.8: Extractions from the Allende and Tagish Lake meteorites, with two profiles from the Nakhla meteorite: the "Free" (aqueous) and "Total" (acid-hydrolyzed) extractions.

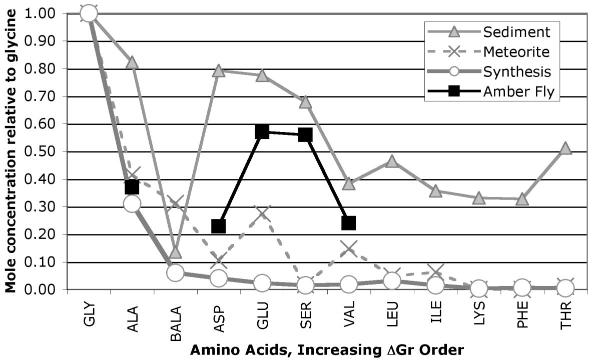


Figure 6.9: Amino acids extracted from an amber-encased fly, compared to the training data set.

scores for both. We believe the extent to which the value is above zero for both measurements is due to the fact that the unreported amino acids VAL, LEU, ILE, LYS, PHE, and THR were set to zero for input into the neural network. Since some of these amino acids have high concentrations in known biotic data, but are always low in the abiotic training set, setting these absent measurements to zero teands to increase the resemblance to the abiotic training set. The generally moderate contributions of ASP, GLU, and SER, and absence of BALA are nonetheless sufficiently distinct from the abiotic base class to generate a value < 0.5.

Ostrom et al. report amino acids in several samples of late cretaceous fossils and bones [72]. Though not protected as well as the amber inclusion in [8], the researchers find enantiomeric and isotopic evidence of persistent amino acids in the fossils they believe may have been protected from racemization by binding to a mineral phase. However, they cannot entirely rule out contamination. As would be expected (whether contaminated or not), our results are consistent with a biotic source for these residues, with data quite similar to the amber inclusion.

6.8 Discussion

The technique described here has shown some promise in generating a discriminator that is nearly independent of knowledge about terrestrial biochemistry, and is a significant improvement on our previous technique. In particular, the confident

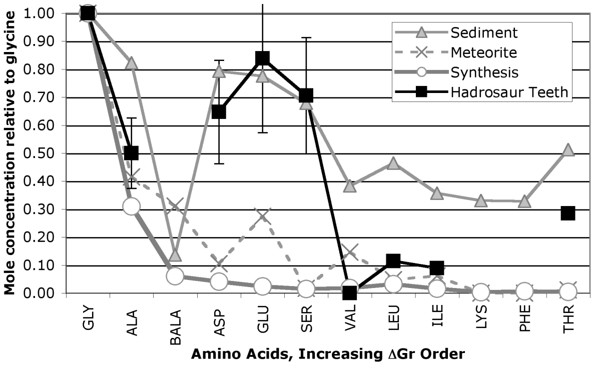


Figure 6.10: Amino acids extracted from hadrosaur teeth, compared to the training data set.

classification of the chondrites Orgueil, Ivuna, and Essebi as outside the known abiotic class is a promising result, since those data are quite distinct from anything in the training set and represent an example of how a truly unknown measurement from a not-yet-characterized chemistry would be handled by this type of system. In this study, we are limited by the available data. Combining extractions from (believed) noncontaminated meteorites and laboratory spark-synthesis experiments, only twenty-nine measurements in available literature quantified a sufficient number of amino acids for inclusion in our training data set $\boldsymbol{P}_{\text{abiotic}}$. Any attempt to use this or a similar technique on an automated mission would need to begin with a very thorough experimental exploration of the amino acid syntheses possible under different conditions.

Others have correctly pointed out that quantifying a dozen amino acids on a distant world is difficult. Here, we use quantified amino acids to demonstrate the technique because they are well-studied and a fair volume of published data exist. However, the method is general and can be extended to other measurements that may be easier to make. For example, the distribution of any particular family of organic chemicals that is detectable in the volatilizable fraction of a soil sample at any particular temperature is likely to be altered noticeably if biota are present that synthesize any of those molecules. Therefore an experiment resembling those used on the Viking landers could be easily constructed using a sample oven and GCMS (gas chromatograph mass spectrometer) without involving the intricacies of amino acid hydrolysis and extraction.

For this strategy to be successful at all, mission designers would, of course, need to select a set of measurements that happened to include monomers used by the putative life form. To the extent that it is applicable, however, known biochemistry leads us to believe that this is not likely to be a significant problem: terrestrial biota use members of nearly every known organic chemical family in a significant way. Each chemical family has functions that others cannot replicate, and known life exploits most of them, suggesting that any other organic life would likely impact at least a few of the chemical concentrations we might choose to characterize.

The presence of biologically synthesized organics in a sample would necessarily alter the total relative molar amounts of carbon, oxygen, nitrogen, and hydrogen in that sample. For example, sterile soils could conceivably contain any of a multitude of minerals, a number of small carbon-containing molecules such as CO, CO_2, CH_4, and limited quantities of small organics like glycine, acetic acid, etc. But, the presence of elevated quantities of any large organic molecules like long-chain carboxylics, larger amino acids, or polycyclic aromatic hydrocarbons could create ratios of C:O, C:H, O:N, etc. that fall outside the range of plausible ratios in the absence of life. So, with a thorough experimental and mathematical analysis of the element ratios of plausible abiotic soils or the volatile fractions thereof, a probe equipped with analytic techniques like the one presented here could detect aberrant (and hence potentially biotic) patterns simply by measuring via mass spectrometry the elemental composition of a vaporized sample.

Part IV

Ongoing Work: The Layered Trophic Residue Biosignature

Part IV consists of a single chapter that discusses ongoing work into another potential universal biosignature, the layered trophic residue biosignature or LTRB. Similar to the MADB, the LTRB looks for life by comparing the concentrations of several compounds. However, the LTRB is concerned with the spatial relationships of active metabolites in samples with physical extent, and how those relationships correlate with the energetic potential of the metabolites as measured by their redox intensity, pε^o.

Chapter 7

The Layered Trophic Residue Biosignature

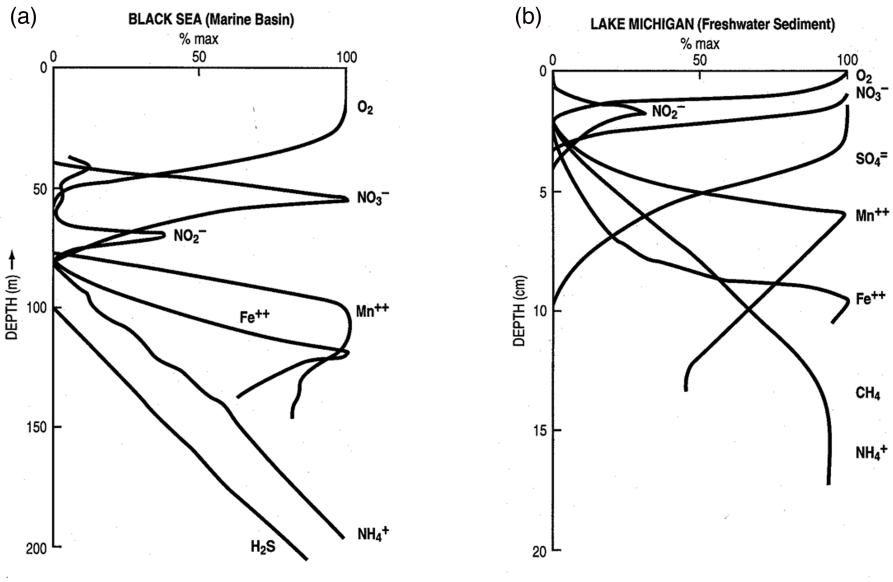


Figure 7.1: Similar layers of dissolved metabolites are found in sediment cores and water columns. Oxic layers appear on top, because oxygen provides the greatest energy potential in respiration. Lower layers are constrained to employ a series of decreasingly effective metabolites.

I propose a second biosignature, the *layered trophic residue biosignature* (LTRB), which I believe represents fundamental processes inherent to life and may therefore be useful for the discovery of extraterrestrial life forms. Work is ongoing to model the evolution and formation of this biosignature in digital life, and to propose methods for automated detection.

It has been observed that communities of microorganisms near phase boundaries (i.e., air to water, or water to sediment) often exist in a layered structure, and that the structure of these communities can be inferred from depth profiles of metabolically active solutes [68, 32]. Figure 7.1, reproduced from [68], shows two examples, remarkable in their similarity, despite a large difference in scale. Figure 7.1a shows the varying concentrations of dissolved oxygen, nitrate, sulfate, ionic manganese and iron, ammonia, and a few others in a 200 m water column in the Black Sea. Figure 7.1b shows a similar profile in the top 20 cm of freshwater sediment at the bottom of Lake Michigan. In both cases, oxygen is present in abundance at the boundary, but rapidly tails off to be replaced in sequence by nitrate, nitrite, and finally, ammonia and reduced iron.

Oxidant	**Redox intensity**	**Energy released**
	pε^o (w)	ΔG^o kJ/equiv. CH_2O
$O_2 \rightarrow H_2O$	13.75	-125
$NO_3^- \rightarrow N_2$	12.65	-119
$MnO_2 \rightarrow Mn^{++}$ (as $MnCO_3$)	8.9	
$NO_3^- \rightarrow NH_4^+$	6.15	-82
$FeOOH \rightarrow Fe^{++}$ (as $FeCO_3$)	-0.8	
$SO_4 \rightarrow HS^-$	-3.75	-25
$CO_2 \rightarrow CH_4$	-4.13	-23

Table 7.1: Several oxidants used as electron-acceptors in metabolism, shown with standard redox intensity. Energy released per equivalent of carbohydrate is shown for those oxidants used in carbohydrate metabolism. Values are all at pH=7 in water, from [86].

If we consider the standard redox intensity, pε^o, of each of these metabolites (Table 7.1), the process that forms this progression of layers becomes clear. Oxygen has the lowest redox potential in this group, and therefore provides the most energy when used as an electron acceptor in metabolism. An aerobic community exists near the boundary of both of these systems, and the oxygen is quickly consumed by these organisms. After the oxygen is gone, however, life persists by using progressively less-energetic metabolites: where aerobic respiration releases 125 kJ/equiv. of carbohydrate, nitrate reduction produces 119kJ/equiv.: more than sufficient for survival, but sub-optimal as compared to aerobic respiration. In some cases, subsequent layers use the byproducts of the layers above — as when sulfide oxidizers use oxygen and HS^- and produce sulfate (100 kJ/equiv.), which is then used in a lower layer of sulfate reduction to oxidize organic carbon (25 kJ/equiv.). This pairing of layers results in an increase and then subsequent decrease with depth of the concentration of SO_4^{2-}, which can be seen in Figure 7.1. Similar processes recycle manganese and iron. The biological strata responsible for these profiles may be communities of different species, or the same species expressing different phenotypes, as some organisms possess the genetic code necessary to survive in more than one chemical environment.

Selection enforces the ordering of the layers. When oxygen is available in abundance, it is maladaptive to manufacture the proteins necessary for the respiration of nitrate; relative to consuming more oxygen, this is a waste of energy. So, aerobic specialists will consume the available oxygen. Specialists which respire other oxidants will be unable to compete with the aerobes and will be forced to occupy a position further from the boundary. An analogous phenomenon occurs for each subsequent trophic layer.

While abiotic processes can certainly form layered structures (e.g., stromatolites), the spatial positioning of the layers *in redox order* indicates that metabolism is active and that selection is enforcing structure. The spatial ordering of chemical strata in redox sequence may therefore qualify as a phenomenon diagnostic for the presence of life. Like the monomer abundance distribution biosignature, the LTRB is agnostic of life's particular biochemistry: the energy available from any particular redox coupling is dictated by thermodynamics. The ordering of preferred resources would be the same for any life form regardless of the chemical nature of its cellular machinery.

This phenomenon can be demonstrated in digital life forms. Avida allows the researcher to specify a number of virtual "resources," the consumption of which can be tied to successful completion of mathematical tasks. Organisms are only rewarded for completing a task if the appropriate resource is available in sufficient quantity. Evolving the code to complete such a task is seen as analogous to evolving the proteins necessary for metabolism of an environmental resource. Others have demonstrated in Avida that the presence of multiple resources leads to speciation when resources are available at depletable levels [15]. Generalists cannot compete with groups of specialists, because they spread their resources too thinly. Smaller groups can consume different resources, thereby more successfully exploiting limited niches. Resources in Avida are, historically, "global," meaning that the same reservoir of resources is available to all cells in the population.

In order to demonstrate the evolution of layered structure, Avida must be reprogrammed to allow resources to take on spatial position within the plane on which avidians live. Resources should be able to flow and diffuse across the grid, and have discrete loci of inflow and outflow. When organisms consume resources, they should consume them locally. Some initial development into this capability has been performed by the Avida development team, but the modelled diffusion physics and software stability are not yet satisfactory. Nonetheless, I have performed some early experimentation modelling the evolution of the LTRB in Avida using these tools.

7.1 Experimental setup

To demonstrate layered communities, I create an Avida environment in which four abstract resources, A, B, C, and D are associated with four boolean tasks: AND, ANDN, ORN, and OR. An avidian which successfully computes the AND task will consume a unit of resource A from its cell on the grid, and be rewarded with an additive bonus to its merit of 1000. (Merit is an avidian's "score"; CPU time is awarded to the organism in proportion to its merit). If there is less than 1.0 units of resource available, then no resource is consumed and no merit reward is given; this is considered equivalent to the manufacture of metabolic enzymes when no appropriate food source is available. The subsequent three resources B, C, and D provide additive merit bonuses of 100, 10, and 1 when the tasks associated with them are completed. Avidians are allowed to complete each task as many as three times per life cycle.

I created a rectangular grid, 20 by 100 cells, and provided a continuous inflow of all four resources into just the top row of 20 cells. Several experiments were performed with the amount of inflow titrated logarithmically between 10 units per update and 1000 units per update. (An update is a unit of time in Avida; replication of an avidian typically takes a few updates). I provided a continuous outflow of all resources, 1.0 units per update, from the bottom row of cells. I also included a fifth, unmetabolizable, resource as a control to show the profile a resource assumes as a result of the inflow, outflow, and diffision parameters. If a resource is not used by the avidians, we would expect its profile to match the control. Initial organisms were generalists, evolved prior to the experiment to include the genetic material necessary to compute all four tasks. I measured the average concentration of the four resources in each row of the grid to see if differential consumption, indicating a layered community structure, had appeared. I also disabled a standard Avida

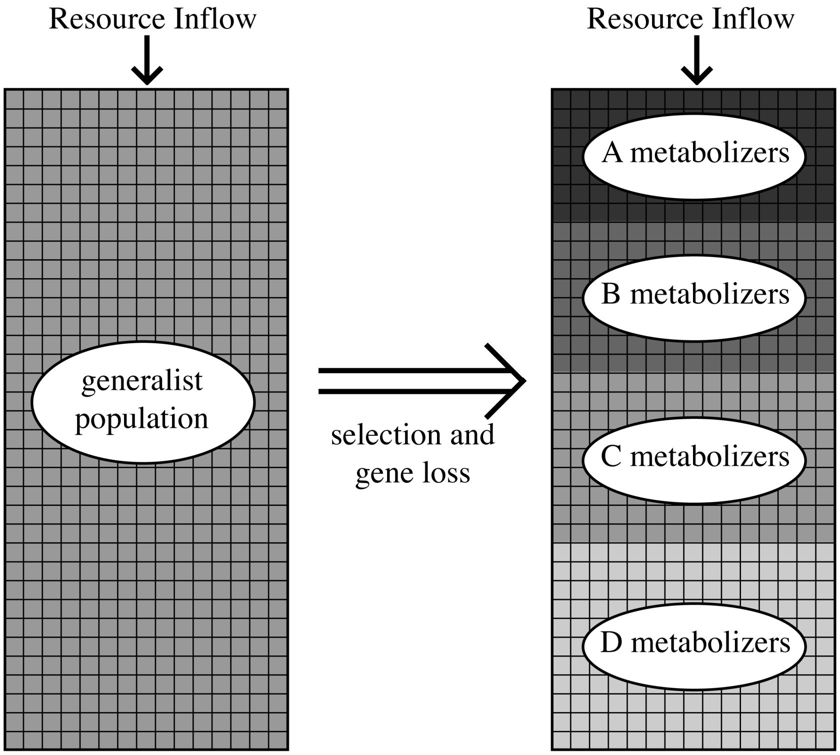


Figure 7.2: A schematic of the Avida environment used to evolve layered community structure. Avidians live on a grid through which resources may diffuse. Four different resources flow continuously into the top row of the grid. The system is initialized with a single genotype that can metabolize all four resources. Selective pressure (optimization for short genomes) is expected to drive the system toward speciation in a layered structure, with upper layers preferring the more energetic resources.

option, "Set base merit proportional to length," which normally allows avidians to increase their genome length without penalty. By disabling it, all organisms are given the same amount of time to reproduce, regardless of length. This provides a strong fitness advantage to genomes that can successfully reduce their length, and therefore, a strong incentive for the organisms to jettison unnecessary genetic material.

7.2 Results

Initial results have been promising. Figure 7.3 shows the concentration profiles of four resources in populations that were evolved in environments with five different resource inflow rates. In all five experiments, resource D is not metabolized at all.

I assume that its metabolic potential is simply too low to make the burden of the genetic code necessary to compute the associated task worthwhile.

Similar to the results of [15], I find that intermediate levels of resource availability are necessary for formation of stable, diverse communitites. At the lowest level of inflow (10 units of resource per update), there are insufficient available resources for organisms to maintain metabolic genes at all. Any attempt to maintain or evolve metabolic genes results in quick depletion of the local resources, leaving those cells carrying useless genetic material. In that experiment, all the resources remain largely untouched. At a slightly higher level of inflow (32 units per update), barely enough resource is available for the organisms to maintain a metabolic gene. In this particular case, they metabolize resource C, consuming all of it in the first three or four rows. I believe that the use of C in particular is precisely because it is the least-energetic of the resources, providing the smallest benefit. If the organisms were to metabolize A or B, they would speed up so much relative to the inflow that they would deplete the resource too quickly. Resource C provides a small bonus that keeps metabolism within the limits where resources will remain available.

When I increase resource inflow to higher levels, we begin to see true layered communities form. At 100 units per update, An active A-metabolizing layer consumes all of that resource within the first four or five rows. A B-metabolizing layer consumes that resource between rows four and ten, and a C-metabolizing resource consumes that resource in approximately rows 10 through 36. That each subsequent layer is slightly deeper than the previous one is an expected result: since the resources provide progressively smaller metabolic benefits, the cell cycle and therefore, resource utilization of each layer is slower, allowing for more organisms to share the resource. When inflow is increased to 316 per update, there is little substantive change to the community except that the B and C layers are thicker. The A layer is actually *thinner*, consuming all of the available A in the top two rows. This probably indicates negative frequency-dependent selection among A-metabolizers in the 100 units per update case. There is not actually enough of resource A to maintain a robust population at that rapid metabolic rate, so the first few layers are a mix of A- and B-metabolizers. Above

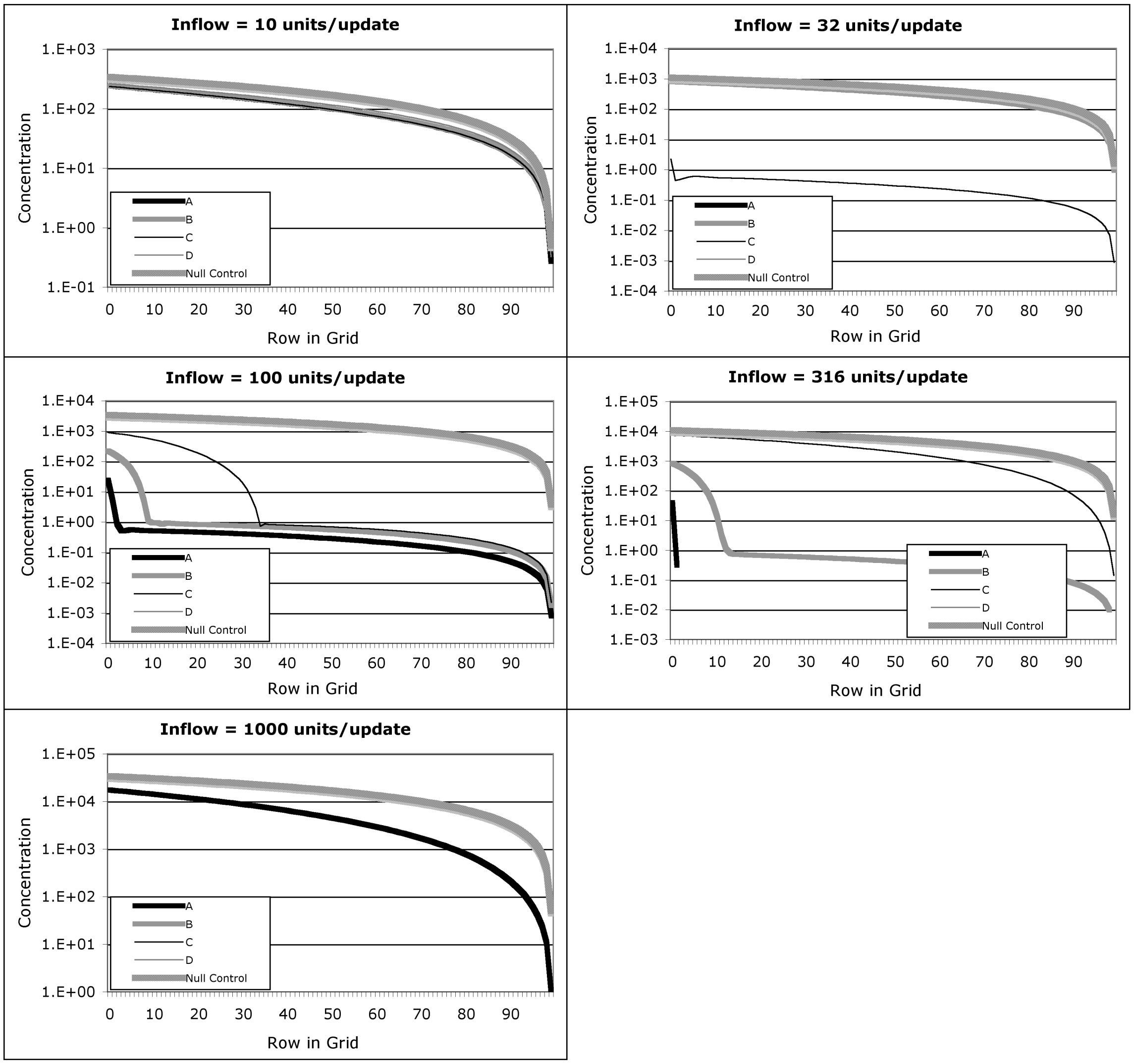


Figure 7.3: Layered consumption of resources in Avida. Profiles of resource concentration in stable, equilibrated evolved populations are shown for five different inflow rates.

a certain threshhold, increases in the population of A-metabolizers reduce the fitness of that genotype as a result of competition for scarce resources.

At the highest tested inflow rate of 1000 units per update, an abundance of resource A is available throughout the population, and all genotypes present metabolize that resource and no other.

In the future of this project, I plan to study the phenomenon in detail, similar to the exhaustive analysis of the MADB presented in Chapters 2 and 3. Parameters of the LTRB I plan to test include:

1. The dependence of layered speciation on the difference in metabolic potential of resources: how much more energetic does A have to be than B before the community will stratify?

2. The dependence on the metabolic cost of genome length: is a penalty for carrying extra genomic material necessary to cause avidians to specialize vie the jettisoning of genes, or will genetic drift alone eliminate unused genes in layers where some resources are unavailable? Can the metabolic cost of genome length be reduced to the point where adaptive radiation will occur starting from a *specialist*, even if some genome length increase is required?

3. Time dependence: if the inflow rate is varying in time, will organisms maintain the genes for metabolizing multiple substrates?

4. Byproducts: demonstrations of lower layers consuming the outputs of upper-layer metabolism.

Bibliography

[1] Abelson, P. H. (1965). Abiogenic synthesis in Martian environment. *Proc. Natl. Acad. Sci. U. S. A.*, *54*, 1490–1494.

[2] Adami, C. (1998). *Introduction to Artificial Life*. New York: Springer.

[3] Adami, C., Ofria, C., & Collier, T. C. (2000). Evolution of biological complexity. *Proc. Natl. Acad. Sci. USA*, *97*, 4463–4468.

[4] Amend, J. P., & Shock, E. L. (1998). Energetics of amino acid synthesis in hydrothermal ecosystems. *Science*, *281*, 1659–1662.

[5] Baath, E., Frostegard, A., & Fritze, H. (1992). Soil bacterial biomass, activity, phospholipid fatty-acid pattern, and pH tolerance in an area polluted with alkaline dust deposition. *Appl. Environ. Microbiol.*, *58*, 4026–4031.

[6] Bada, J. L., & McDonald, G. D. (1995). Amino-acid racemization on Mars – implications for the preservation of biomolecules from an extinct Martian biota. *Icarus*, *114*, 139–143.

[7] Bada, J. L., & McDonald, G. D. (1996). Detecting amino acids on Mars. *Anal. Chem.*, *68*, A668–A673.

[8] Bada, J. L., Wang, X. S., Poinar, H. N., Paabo, S., & Poinar, G. O. (1994). Amino-acid racemization in amber-entombed insects – implications for DNA preservation. *Geochim. Cosmochim. Acta.*, *58*, 3131–3135.

[9] Bell, G. (2001). Neutral macroecology. *Science*, *293*, 2413–2418.

[10] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

[11] Botta, O., & Bada, J. L. (2002). Extraterrestrial organic compounds in meteorites. *Surv. Geophys.*, *23*, 411–467.

[12] Botta, O., Glavin, D. P., Kminek, G., & Bada, J. L. (2002). Relative amino acid concentrations as a signature for parent body processes of carbonaceous chondrites. *Origins Life Evol. Biosph.*, *32*, 143–163.

[13] Brooks, D. J., Fresco, J. R., Lesk, A. M., & Singh, M. (2002). Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.*, *19*, 1645–1655.

[14] Centler, F., Dittrich, P., Ku, L., Matsumaru, N., Pfaffmann, J., & 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 of Artificial Life ECAL*, Lecture Notes in Artificial Intelligence, (31–40). Berlin: Springer.

[15] Chow, S. S., Wilke, C. O., Ofria, C., Lenski, R. E., & Adami, C. (2004). Adaptive radiation from resource competition in digital organisms. *Science*, *305*, 84–86.

[16] Colombo, J. C., Silverberg, N., & Gearing, J. N. (1996). Lipid biogeochemistry in the Laurentian Trough .1. fatty acids, sterols and aliphatic hydrocarbons in rapidly settling particles. *Org. Geochem.*, *25*, 211–225.

[17] Colombo, J. C., Silverberg, N., & Gearing, J. N. (1998). Amino acid biogeochemistry in the Laurentian Trough: vertical fluxes and individual reactivity during early diagenesis. *Org. Geochem.*, *29*, 933–945.

[18] Cowie, G. L., Hedges, J. I., & Calvert, S. E. (1992). Sources and relative reactivities of amino-acids, neutral sugars, and lignin in an intermittently anoxic marine environment. *Geochim. Cosmochim. Acta.*, *56*, 1963–1978.

[19] Cronin, J. R., Cooper, G. W., & Pizzarello, S. (1994). Characteristics and formation of amino-acids and hydroxy-acids of the Murchison meteorite. In *Life Sciences and Space Research XXV (4)*, vol. 15 of *Advances in Space Research*, (91–97). Oxford: Pergamon Press Ltd.

[20] Cronin, J. R., Gandy, W. E., & Pizzarello, S. (1981). Amino-acids of the Murchison meteorite .1. 6 carbon acyclic primary alpha-amino alkanoic acids. *J. Mol. Evol.*, *17*, 265–272.

[21] Cronin, J. R., & Moore, C. B. (1971). Amino acids analyses of the Murchison, Murray, and Allende carbonaceous chondrites. *Science*, *172*, 1327–1329.

[22] Cronin, J. R., & Moore, C. B. (1976). Amino-acids of Nogoya and Mokoia carbonaceous chondrites. *Geochim. Cosmochim. Acta.*, *40*, 853–857.

[23] Cronin, J. R., Moore, C. B., & Pizzarello, S. (1980). Amino-acids in six CM2 chondrites. *Meteoritics*, *15*, 277–278.

[24] Cronin, J. R., & Pizzarello, S. (1983). Amino acids in meteorites. *Adv. Space Res.*, *3*, 5–18.

[25] Cronin, J. R., & Pizzarello, S. (1997). Enantiomeric excesses in meteorite amino acids. *Abstracts of Papers of the American Chemical Society*, *213*, 215–GEOC.

[26] Cronin, J. R., Pizzarello, S., & Cruikshank, D. P. (1988). Organic matter in carbonaceous chondrite, planetary satellites, asteroids and comets. In J. F. Kerridge, & M. S. Matthews (Eds.) *Meteorites and the Early Solar System*, (114–143). Tucson: University of Arizona Press.

[27] Cronin, J. R., Pizzarello, S., & Moore, C. B. (1979). Amino-acids in an Antarctic carbonaceous chondrite. *Science*, *206*, 335–337.

[28] Dauwe, B., & Middelburg, J. J. (1998). Amino acids and hexosamines as indicators of organic matter degradation state in North Sea sediments. *Limnol. Oceanogr.*, *43*, 782–798.

[29] Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure*, vol. 5. Washington, D.C.: National Biomedical Research Foundation.

[30] Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure, Supplement 3*, vol. 5. National Biomedical Research Foundation.

[31] Deamer, D. W. (1999). How did it all begin? The self-assembly of organic molecules and the origin of cellular life. In J. Scotchmoor, & D. A. Springer (Eds.) *Evolution: Investigating the Evidence*, vol. 9. Knoxville: The Paleontological Society.

[32] D'Hondt, S., Jorgensen, B. B., Miller, D. J., Batzke, A., Blake, R., Cragg, B. A., Cypionka, H., Dickens, G. R., Ferdelman, T., Hinrichs, K. U., Holm, N. G., Mitterer, R., Spivack, A., Wang, G., Bekins, B., Engelen, B., Ford, K., Gettemy, G., Rutherford, S. D., Sass, H., Skilbeck, C. G., Aiello, I. W., Guerin, G., House, C. H., Inagaki, F., Meister, P., Naehr, T., Niitsuma, S., Parkes, R. J., Schippers, A., Smith, D. C., Teske, A., Wiegel, J., Padilla, C. N., & Acosta, J. L. (2004). Distributions of microbial activities in deep subseafloor sediments. *Science*, *306*, 2216–2221.

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

[34] Dorn, E. D., Nealson, K. H., & Adami, C. (2005). Monomer abundance patterns as a universal biosignature: Examples from terrestrial and artificial life. *J. Mol. Evol., submitted*.

[35] Engel, M. H., & Macko, S. A. (1997). Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature*, *389*, 265–268.

[36] Engel, M. H., Macko, S. A., & Silfer, J. A. (1990). Carbon isotope composition of individual amino-acids in the Murchison meteorite. *Nature*, *348*, 47–49.

[37] Engel, M. H., & Nagy, B. (1982). Distribution and enantiomeric composition of amino-acids in the Murchison meteorite. *Nature*, *296*, 837–840.

[38] Gebicki, J. M., & Hicks, M. (1976). Preparation and properties of vesicles enclosed by fatty acid membranes. *Chem Phys Lipids*, *16*, 142–160.

[39] Glavin, D. P., Bada, J. L., Brinton, K. L. F., & McDonald, G. D. (1999). Amino acids in the Martian meteorite Nakhla. *Proc. Natl. Acad. Sci. U. S. A.*, *96*, 8835–8838.

[40] Gorban, A. N., Zinovyev, A. Y., & Popova, T. G. (2003). Seven clusters in genomic triplet distributions. *In Silico Biol.*, *3*, 471–482.

[41] Hargreaves, W. R., & Deamer, D. W. (1978). Liposomes from ionic, single-chain amphiphiles. *Biochemistry*, *17*.

[42] Hedges, J. I., Mayorga, E., Tsamakis, E., McClain, M. E., Aufdenkampe, A., Quay, P., Richey, J. E., Benner, R., Opsahl, S., Black, B., Pimentel, T., Quintanilla, J., & Maurice, L. (2000). Organic matter in bolivian tributaries of the amazon river: A comparison to the lower mainstream. *Limnol. Oceanogr.*, *45*, 1449–1466.

[43] Hedges, J. I., & Oades, J. M. (1997). Comparative organic geochemistries of soils and marine sediments. *Org. Geochem.*, *27*, 319–361.

[44] Horsfall, I. M., & Wolff, G. A. (1997). Hydrolysable amino acids in sediments from the porcupine abyssal plain, northeast Atlantic Ocean. *Org. Geochem.*, *26*, 311–320.

[45] Keil, R. G., & Fogel, M. L. (2001). Reworking of amino acid in marine sediments: Stable carbon isotopic composition of amino acids in sediments along the washington coast. *Limnol. Oceanogr.*, *46*, 14–23.

[46] Keil, R. G., Tsamakis, E., Giddings, J. C., & Hedges, J. I. (1998). Biochemical distributions (amino acids, neutral sugars, and lignin phenols) among size-classes

of modern marine sediments from the Washington coast. *Geochim. Cosmochim. Acta.*, *62*, 1347–1364.

[47] Khare, B. N., Sagan, C., Ogino, H., Nagy, B., Er, C., Schram, K. H., & Arakawa, E. T. (1986). Amino-acids derived from Titan tholins. *Icarus*, *68*, 176–184.

[48] Kielland, K. (1995). Landscape patterns of free amino acids in arctic tundra soils. *Biogeochemistry*, *31*, 85–98.

[49] Kminek, G., Botta, O., Glavin, D. P., & Bada, J. L. (2002). Amino acids in the Tagish Lake meteorite. *Meteorit. Planet. Sci.*, *37*, 697–702.

[50] Kvenvolden, K. A., Lawless, J., Pering, K., Peterson, E., Flores, J., Ponnamperuma, C., Kaplan, I. R., & Moore, C. (1970). Evidence for extraterrestrial amino-acids and hydrocarbons in Murchison meteorite. *Nature*, *228*, 923–926.

[51] Kvenvolden, K. A., Lawless, J. G., & Ponnamperuma, C. (1971). Nonprotein amino acids in Murchison meteorite. *Proc. Natl. Acad. Sci. U. S. A.*, *68*, 486–490.

[52] Lawless, J. G., & Yuen, G. U. (1979). Quantification of monocarboxylic acids in the Murchison carbonaceous meteorite. *Nature*, *282*, 396–398.

[53] Lenski, R. E., Ofria, C., Collier, T. C., & Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, *400*, 661–664.

[54] Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*, 139–144.

[55] Lerner, N. R., Peterson, E., & Chang, S. (1993). The Strecker synthesis as a source of amino-acids in carbonaceous chondrites – deuterium retention during synthesis. *Geochim. Cosmochim. Acta.*, *57*, 4713–4723.

[56] MacDermott, A. J., Barron, L. D., Brack, A., Buhse, T., Drake, A. F., Emery, R., Gottarelli, G., Greenberg, J. M., Haberle, R., Hegstrom, R. A., Hobbs,

K., Kondepudi, D. K., McKay, C., Moorbath, S., Raulin, F., Sandford, M., Schwartzman, D. W., Thiemann, W. H. P., Tranter, G. E., & Zarnecki, J. C. (1996). Homochirality as the signature of life: The SETH cigar. *Planetary and Space Science*, *44*, 1441–1446.

[57] McDonald, G. D., Khare, B. N., Thompson, W. R., & Sagan, C. (1991). CH4/NH3/H2O spark tholin – chemical-analysis and interaction with Jovian aqueous clouds. *Icarus*, *94*, 354–367.

[58] McDonald, G. D., Storrie-Lombardi, M. C., & Nealson, K. H. (1999). Principal component analysis for biosignature detection in extraterrestrial samples. *Abstracts of Papers of the American Chemical Society*, *217*, U846–U846.

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

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

[61] Miller, S. L. (1955). Production of some organic compounds under possible primitive Earth conditions. *J. Am. Chem. Soc.*, *77*, 2351–2361.

[62] Miller, S. L. (1957). The formation of organic compounds on the primitive Earth. *Annals of the New York Academy of Sciences*, *69*, 260–275.

[63] Miller, S. L., & Urey, H. C. (1959). Organic compound synthesis on the primitive Earth. *Science*, *130*, 245–251.

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

[65] Nagy, B., & Bitz, S. M. C. (1963). Long-chain fatty acids from the Orgueil meteorite. *Arch. Biochem. Biophys.*, *101*, 240.

[66] Naraoka, H., Shimoyama, A., & Harada, K. (1996). Molecular distribution of monocarboxylic acids in Asuka carbonaceous chondrites from Antarctica. *Origins Life Evol. Biosph.*, *29*, 187–201.

[67] Nealson, K. H. (2002). The limits of life on Earth and searching for life on Mars. *J. Geophys. Res. Planets*, *102*, 23675–23686.

[68] Nealson, K. H., & Stahl, D. A. (1997). Microorganisms and biogeochemical cycles: What can we learn from layered microbial communities? In *Geomicrobiology: Interactions between Microbes and Minerals*, vol. 35 of *Reviews in Mineralogy*, (5–34). Washington: Mineralogical Society of America.

[69] Ofria, C., Adami, C., & Collier, T. C. (2002). Design of evolvable computer languages. *IEEE Trans. on Evol. Comp.*, *6*, 420–424.

[70] Ofria, C., Adami, C., & Collier, T. C. (2003). Selective pressures on genomes in molecular evolution. *J. Theor. Biol.*, *222*, 477–483.

[71] Ofria, C., & Wilke, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, *10*, 191–229.

[72] Ostrom, P. H., Macko, S. A., Engel, M. H., Silfer, J. A., & Russell, D. (1990). Geochemical characterization of high-molecular-weight material isolated from late cretaceous fossils. *Org. Geochem.*, *16*, 1139–1144.

[73] Ray, T. S. (1992). An approach to the synthesis of life. In C. G. Langton, J. D. Farmer, & S. Rasmussen (Eds.) *Artificial Life II*, (371–408). Redwood City: Addison-Wesley.

[74] Ring, D., Wolman, Y., Miller, S. L., & .N, F. (1972). Prebiotic synthesis of hydrophobic and protein amino-acids. *Proc. Natl. Acad. Sci. U. S. A.*, *69*, 765–768.

[75] Rushdi, A. I., & T., S. B. R. (2001). Lipid formation by aqueous Fischer-Tropsch-type synthesis over a temperature range of 100 to 400^{o}C. *Origins Life Evol. Biosph.*, *31*, 103–118.

[76] Schlesinger, G., & Miller, S. L. (1983). Prebiotic synthesis in atmospheres containing CH4, CO, and CO2. 1. amino-acids. *J. Mol. Evol.*, *19*, 376–382.

[77] Schlesinger, G., & Miller, S. L. (1986). Prebiotic syntheses of pantoic acid and the other components of coenzyme-a. *Origins Life Evol. Biosph.*, *16*, 307–307.

[78] Schultes, E. A., Hraber, P. T., & LaBean, T. H. (1997). Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA*, *3*, 792–806.

[79] Schultes, E. A., Hraber, P. T., & LaBean, T. H. (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, *49*, 76–83.

[80] Shimoyama, A., Ikeda, H., Nomoto, S., & Harada, K. (1994). Formation of carboxylic-acids from elemental carbon and water by arc-discharge experiments. *Bull. Chem. Soc. Jpn.*, *67*, 257–259.

[81] Shimoyama, A., Komiya, M., & Harada, K. (1991). Low-molecular-weight monocarboxylic acids and gamma-lactones in neogene sediments of the Shinjo Basin. *Geochem. J.*, *25*, 421–428.

[82] Shimoyama, A., Naraoka, H., Yamamoto, H., & Harada, K. (1986). Carboxylic acids in the Yamato-791198 carbonaceous chondrites from antarctica. *Chem. Lett.*, *15*, 1561–1564.

[83] Shimoyama, A., Ponnamperuma, C., & Yanai, K. (1979). Amino-acids in the Yamato carbonaceous chondrite from Antarctica. *Nature*, *282*, 394–396.

[84] Smith, J. M. (1992). Evolutionary biology - byte-sized evolution. *Nature*, *355*, 772–773.

[85] Storrie-Lombardi, M. C., Irwin, M., & von Hippel, T. (1994). Spectral classification with principal component analysis and artificial neural networks. *Vistas in Astronomy*, *38*, 331–340.

[86] Stumm, W., & Morgan, J. J. (1996). *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*. New York: John Wiley and Sons, third edn.

[87] Sugisaki, R., & Mimura, K. (1994). Mantle hydrocarbons - abiotic or biotic. *Geochim. Cosmochim. Acta.*, *58*, 2527–2542.

[88] Sundh, I., Nilsson, M., & Borga, P. (1997). Variation in microbial community structure in two boreal peatlands as determined by analysis of phospholipid fatty acid profiles. *Appl. Environ. Microbiol.*, *63*, 1476–1482.

[89] Takano, Y., Ohashi, A., Kaneko, T., & Kobayashi, K. (2004). Abiotic synthesis of high-molecular-weight organics from an inorganic gas mixture of carbon monoxide, ammonia, and water by 3 mev proton irradiation. *Appl. Phys. Lett.*, *84*, 1410–1412.

[90] Wakeham, S. G. (1999). Monocarboxylic, dicarboxylic and hydroxy acids released by sequential treatments of suspended particles and sediments of the black sea. *Org. Geochem.*, *30*, 1059–1074.

[91] Wang, X. S., Poinar, H. N., Poinar, G. O., & Bada, J. L. (1995). Amino acids in the amber matrix and in entombed insects. In J. Anderson, & J. C. Crelling (Eds.) *Amber, Resinite, and Fossil Resins*, vol. 617 of *ACS Symposium Series*, (255–262). Washington: American Chemical Society.

[92] White, D. C., Ringelberg, D. B., Macnaughton, S. J., Alugupalli, S., & Schram, D. (1997). Signature lipid biomarker analysis for quantitative assessment in situ of environmental microbial ecology. In *Molecular Markers in Environmental Geochemistry*, vol. 671 of *ACS Symposium Series*, (22–34). American Chemical Society.

[93] Wilke, C. O., & Adami, C. (2002). The biology of digital organisms. *Trends Ecol. Evol.*, *17*, 528–532.

[94] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, *412*, 331–333.

[95] Wolman, Y., Haverlan, W. J., & Miller, S. L. (1972). Non-protein amino acids from spark discharges and their comparison with Murchison meteorite amino acids. *Proc. Natl. Acad. Sci. U. S. A.*, *69*, 809–811.

[96] Yedid, G., & Bell, G. (2001). Microevolution in an electronic microcosm. *Am Nat*, *157*, 465–487.

[97] Yuen, G. U., Lawless, J. G., & Edelson, E. H. (1981). Quantification of monocarboxylic acids from a spark discharge synthesis. *J. Mol. Evol.*, *17*, 43–47.

[98] Zelles, L., & Bai, Q. Y. (1994). Fatty-acid patterns of phospholipids and lipopolysaccharides in environmental samples. *Chemosphere*, *28*, 391–411.

Appendix A

Description of Electronic Files

The electronic edition of this thesis includes a number of auxiliary digital files, listed and described below. If you are reading the paper edition of this thesis, you will need to access the Caltech electronic thesis repository at http://etd.caltech.edu/ if you wish to retrieve any of these files.

- **biosignature_movie_1.mpg** (987 KB) — An MPEG format movie showing the development of the MAB in Avida as described in 2. This movie shows the relative abundances of 28 Avida instructions in a single typical evolutionary run. This run was seeded with the standard progenitor (listed in Table 2.4.2.1) and run for 200,000 updates with copy mutation probability μ=0.01 randomizing mutations per execution of the COPY instruction. The red curve shows the initial distribution of instructions present in the progenitor population, while the blue curve shows the actual distribution of instructions at any particular point in time. The distribution was sampled every 100 updates. The run illustrated here is one of twenty-five runs used to generate the "0.01" curve in Figure 2.4a.

- **biosignature_movie_2.mpg** (1.3 MB) — An MPEG-format movie that illustrates the experiment shown in Figure 2.5. Figure 2.5. shows how the concentration of six instructions varies over time as the point mutation rate is ramped down from lethal to non-lethal levels and back again in an environment continually re-seeded with viable organisms. This movie shows the distribution curve of all 28 instructions as it varies over the same time period.

  As can be clearly seen in the animation, as long as the point mutation rate remains too high, the organisms cannot survive and the distribution of instructions in the environment is even, reflecting the random nature of the mutation function being continually applied. Once the environment becomes nonlethal, a population explosion quickly establishes a biosignature. As the population adapts to its environment the pattern rapidly comes to match the familiar “adapted” biosignature shown in Figure 2.4.

  Once the point mutation rate is again increased to inhospitable levels a population crash ensues and the biosignature disappears as abiotic processes again dominate.

- **LTRB_movie.mpg** (915 KB) — An MPEG-format movie demonstrating the development of the LTRB effect as described in Chapter 7. Three columns show two-dimensional maps of the concentration of resources A, B, and C in a narrow column containing an avida population, while a fourth plots a vertical profile of the average concentration of the three resources. As the population speciates, the most valuable resource, A, is rapidly consumed in the uppermost layers, while the less-energetic resources diffuse down to lower layers containing other species.

- **avida_1.6ED1.tgz** (883 KB) — A zipped tar archive including both the source code and an x86 Linux binary of Avida version 1.6ED1, the customized version of Avida used to generate the results described in Chapter 2.

- **avida_1.6ED2.tgz** (1.9 MB) — A zipped tar archive including both the source code and an x86 Linux binary of Avida version 1.6ED2, the customized version of Avida used to generate the results described in Chapters 3 and 4.

- **avida_2.0b7.tgz** (3.3 MB) — A zipped tar archive including both the source code and an x86 Linux binary of Avida version 2.0b7. This standard release version of Avida was used to generate the results described in Chapter 7.

- **config_chapter2A.tgz** (8 KB) — A zipped tar archive including all of the configuration files necessary for reproducing the Avida experiments described in Figure 2.4.

- **config_chapter2B.tgz** (7 KB) — A zipped tar archive including all of the configuration files necessary for reproducing the Avida experiments described in Figure 2.5.

- **config_chapter3.tgz** (11 KB) — A zipped tar archive including all of the configuration files necessary for reproducing the Avida experiments described in Chapter 3.

- **config_chapter4.tgz** (12 KB) — A zipped tar archive including all of the configuration files necessary for reproducing the Avida experiments described in Chapter 4.

- **AminoAcids.xls** (1.1 MB) — A Microsoft Excel workbook containing the collected amino acid data used in Chapters 2, 5, and 6. This is the most recent version and includes more recent data than that used in Chapter 5.

- **CarboxylicAcids.xls** (262 KB) — A Microsoft Excel workbook containing the collected carboxylic acid data used in Chapter 2.