Surveying the sequence diversity of model prebiotic peptides by mass spectrometry

Jay G. Forsythe, Anton S. Petrov, W. Calvin Miller, Sheng-Sheng Yu, Ramanarayanan Krishnamurthy, Martha A. Grover, Nicholas V. Hud, and Facundo M. Fernández

*School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332-0400; †School of Physics, Georgia Institute of Technology, Atlanta, GA 30332-0430; ‡School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0106; and †Department of Chemistry, The Scripps Research Institute, La Jolla, CA 92037

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved August 7, 2017 (received for review June 29, 2017)

The rise of peptides with secondary structures and functions would have been a key step in the chemical evolution which led to life. As with modern biology, amino acid sequence would have been a primary determinant of peptide structure and activity in an origins-of-life scenario. It is a commonly held hypothesis that unique functional sequences would have emerged from a diverse soup of proto-peptides, yet there is a lack of experimental data in support of this. Whereas the majority of studies in the field focus on peptides containing only one or two types of amino acids, here we used modern mass spectrometry (MS)-based techniques to separate and quantify these peptides by mass spectrometry. Successful depsipeptide formation with various amino acid and hydroxy acid monomers was confirmed. At drying temperatures of 55–65 °C, depsipeptide chains contained predominantly ester bonds, whereas amide bonds became more prevalent at drying temperatures above 65 °C. Most reactions tested had an initial pH of ~3 due to the hydroxy acid monomers present, but depsipeptide formation also occurred at pH values of 5, 7, and 9. Successful depsipeptide formation with various amino acid and hydroxy acid monomers was achieved, including glycine, glycolic acid, t-alanine, n-alanine, l-lactic acid, l-leucine, and l-serine.

Motivated by our discovery that ester–amide exchange reactions can produce mixed depsipeptides capable of chemical evolution, we sought to explore the sequence diversity of these species when more plausible mixtures of amino acid and hydroxy acid monomers were subjected to repeated cycles. As with traditional proteomics, these reactions facilitated by our discovery that ester–amide exchange reactions can produce mixed depsipeptides capable of chemical evolution, we sought to explore the sequence diversity of these species when more plausible mixtures of amino acid and hydroxy acid monomers were subjected to repeated cycles. As with traditional proteomics.

Significance

Peptides and proteins are essential for life as we know it, and likely played a critical role in the origins of life as well. In recent years, much progress has been made in understanding plausible routes from amino acids to peptides. However, little is known about the diversity of sequences that could have been produced by abiogenic condensation reactions on the prebiotic earth. In this study, multidimensional separations were coupled with mass spectrometry to detect and sequence mixtures of model proto-peptides. It was observed that, starting with a few monomers, proto-peptide diversity increased rapidly following cycling. Experimental proto-peptide sequences were compared with theoretically random sequences, revealing a high sequence diversity of plausible monomer combinations.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All depsipeptide sequences reported in this paper can be accessed in Dataset S1.

1Present address: Department of Chemistry and Biochemistry, College of Charleston, Charleston, SC 29424.
2Present address: Master’s Industrial Internship Program, University of Oregon, Eugene, OR 97403.
3To whom correspondence should be addressed. Email: facundo.fernandez@chemistry.gatech.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1711631114/DCSupplemental.

E7652–E7659 | PNAS | Published online August 28, 2017

www.pnas.org/cgi/doi/10.1073/pnas.1711631114
research, mass spectrometry (MS) was the logical choice for the investigation of depsipeptide primary structure (i.e., monomer sequence) in the complex reaction mixtures. Significant developments in MS and ancillary technologies (15, 16) have dramatically advanced the ability to deeply characterize biopolymers and proteomes, with MS-based methods enabling the detection of thousands of unique peptides and their modifications in a single chromatographic separation (17–19).

Advances in proteomics, however, have not yet translated to the study of complex proto-peptide mixtures in an origins-of-life scenario. One of the most daunting challenges associated with this question is the inherent vastness of proto-peptide sequence space, which may be many orders of magnitude greater than that of peptides encoded by genes in living cells (20). If a model prebiotic pool of only 10 amino acids and 10 hydroxy acids is considered, and equal monomer reactivity is assumed, then the potential number of depsipeptide sequences with length 10 would be \((10 + 10)^{10}\), or \(~10^{15}\). As a result of this staggering number, progress on the characterization of plausibly prebiotically assembled amino acid polymers (1, 13, 21–25), abiogenic mechanisms for peptide self-assembly (10, 26–30), and the evolution of early ribosomes (31, 32) has not been paralleled by a systems-level understanding of proto-peptide diversity.

Here, a peptidomic workflow was developed for mapping the sequence diversity of depsipeptide libraries generated from combinations of plausibly prebiotic monomers. Because depsipeptides are not found in current protein databases, an in-house algorithm was developed to generate custom in silico sequences. Processed experimental data were matched with theoretical sequences containing hydroxy acids and amino acids via accurate mass and tandem mass spectral similarity searches. Using this workflow, hundreds of depsipeptide sequences were surveyed after only 4 d of cycling. Experimentally observed depsipeptide sequences were compared with the theoretically possible permutations using homology diagrams in order to investigate the sequence space of plausible, de novo-synthesized proto-peptides in an origins-of-life context.

**"Proto-Peptidomics" Workflow for the Analysis of Depsipeptide Mixtures**

The primary justification for using a peptidomics-inspired methodology in this work stems from the chemical similarities between peptides and depsipeptides, as demonstrated by several studies (33–36). In designating monomers and/or residues, a similar nomenclature system to that of Deechongkit et al. (34) was used, where an amino acid residue is capitalized (e.g., glycine = G) and its corresponding hydroxy acid residue is lowercase (e.g., glycolic acid = g). A general overview of the experimental workflow is provided in Fig. 1: detailed methods and raw data are provided in SI Appendix. Three prebiotic amino acids, glycine (G), L-alanine (A), and L-leucine (L), were mixed with one hydroxy acid: glycolic acid (g), L-lactic acid (a), or L-malic acid (d). Each mixture contained one hydroxy acid and three amino acids (g+A+L, a+G+A+L, and d+G+A+L) and was subjected to one, two, or three 85 °C-dry/65 °C-wet cycles chosen to model a daily evaporating pool environment on the early Earth (12). The resulting nine depsipeptide libraries were examined by ultraperformance liquid chromatography (UPLC) and traveling-wave IM (TWIM) before data-dependent MS/MS acquisition. To evaluate the performance of the TWIM detector, a protocol was established where an amino acid/monomer mixture was polymerized (85 °C dry-down, 18 h, open vial) and then hydrolyzed (65 °C aqueous, 6 h, closed vial) conditions. Each cycle consisted of one polymerization and one hydrolysis step over a period of 24 h.
liquid chromatography–ion mobility–tandem mass spectrometry (UPLC-IM-MS/MS) in data-dependent acquisition (DDA) mode. Similar to traditional proteomics experiments, experimental and theoretical sequences were matched following a ranking score (Fig. 1C). Sequence matching was achieved by the development of a script that generated accurate masses and fragment ion databases for all possible compositions and oligomer lengths within the experimental mass-to-charge ratio (m/z) range observed. Databases were then uploaded into a software package for comparison with experimental data. All sequence identifications were manually validated.

Example UPLC-IM and MS/MS data are shown in Fig. 1D and E, respectively. As expected, the datasets were highly complex, even after one dry–wet cycle. When depsipeptides eluted from the UPLC column and the precursor ions were selected for MS/MS activation, their fragmentation patterns were generally consistent with those for biological peptides (37, 38). Therefore, standard peptide fragmentation nomenclature (39) was used (SI Appendix, Figs. S1 and S2).

### Mapping Depsipeptide Chemical Space

Across all cycles, a total of 467 unique features, or unique m/z, UPLC retention time (R_t), and IM drift time (d_t) combinations were observed in the glycolic acid (g+G+A+L) dataset, 624 features were observed in the lactic acid (a+G+A+L) dataset, and 426 features were observed in the malic acid (d+G+A+L) dataset (Table 1).

The total number of depsipeptide compositions (e.g., 2a+1G+1A) confirmed by accurate mass was 176 for the g+G+A+L dataset, 239 for the a+G+A+L dataset, and 45 for the d+G+A+L dataset. Three-dimensional maps (m/z, R_t, d_t) of confirmed depsipeptide compositions are shown in Fig. 2A–C. All depsipeptides contained at least one hydroxy acid residue (at the “N terminus”) and at least one amino acid. Features that tentatively matched depsipeptide masses but were not mass-selected for MS/MS during DDA, or were not abundant enough to generate quality MS/MS spectra, were excluded from these plots. Overall, the results presented in Fig. 2 revealed highly complex mixtures, with hundreds of detected species even after a few environmental cycles.

An interesting observation from the data in Fig. 2A–C is that malic-acid-containing depsipeptides were much fewer in number than glycolic acid or lactic-acid-containing depsipeptides. Although the overall number of features observed in the d+G+A+L dataset was similar to the g+G+A+L and a+G+A+L datasets, the number of confirmed linear depsipeptide sequences was significantly lower. We initially hypothesized that malic-acid-based depsipeptides would be numerous, as malic acid (d) is

### Table 1. Experimentally observed and theoretically possible depsipeptides, with respect to oligomer length

<table>
<thead>
<tr>
<th>Oligomer Length</th>
<th>Effective Length, n</th>
<th>No. of Unique Composition Matches</th>
<th>Unique Sequences Confirmed by MS/MS</th>
<th>Theor. Sequence Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g+G+A+L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>20</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>35</td>
<td>37</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>41</td>
<td>73</td>
<td>256</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>44</td>
<td>91</td>
<td>1,024</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>22</td>
<td>72</td>
<td>4,096</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>10</td>
<td>33</td>
<td>16,384</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>467</td>
<td>176</td>
<td>317</td>
</tr>
<tr>
<td>(a+G+A+L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>42</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>52</td>
<td>37</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>59</td>
<td>83</td>
<td>256</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>55</td>
<td>110</td>
<td>1,024</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>18</td>
<td>67</td>
<td>4,096</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16,384</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>624</td>
<td>239</td>
<td>311</td>
</tr>
<tr>
<td>(d+G+A+L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>20</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
<td>14</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>256</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1,024</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4,096</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16,384</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>426</td>
<td>45</td>
<td>28</td>
</tr>
</tbody>
</table>

*Effective depsipeptide length, not accounting for the first hydroxy acid residue. This corresponds to the number of variable monomers per oligomer.

†Number of unique m/z, retention time, and IM arrival-time combinations detected by commercial MS software after chromatographic alignment.

‡Mass tolerance ±15 ppm.

§Unique sequences determined from MS/MS spectra (manually confirmed). For short oligomers, the same sequence was often observed at multiple retention times due to in-source fragmentation of longer, ester-rich depsipeptides.
similar to aspartic acid (D) used to form Fox’s proteinoids. The carboxylic acid moiety of the malic acid side chain could be capable of polymer branching (Fig. 1A), and its polarity was expected to improve oligomer solubility characteristics. Therefore, it was somewhat surprising to observe that malic acid generated the fewest detectable depsipeptides of the three hydroxy acids investigated.

Cyclic depsipeptides lose an additional water molecule when formed, and thus have a mass 18.0106 Da lower than their corresponding linear depsipeptides. In our initial data analysis, water-loss species were not included in the m/z databases, as they can also occur through unavoidable in-source collision-induced dissociation processes of linear species. Manual reinvestigation of the d+G+A+L dataset, however, indicated that malic acid and leucine cyclic species (cyclo-dL)—isobaric with water-loss fragments, but with unique R<sub>t</sub> and d<sub>4</sub> values—were indeed present in moderate abundance (SI Appendix, Figs. S3–S5). We hypothesize steric hindrance from bulky leucine side chains in cyclic dL products restricted water access to ester bonds and, in turn, led to reduced hydrolysis rates of those species back to the linear form. This could result in a reduced amount of reactive groups available to form new depsipeptides in the d+G+A+L system, potentially similar to diketopiperazine formation in amino-acid-only experiments. Because the diversity of depsipeptide mixtures is dependent upon not only the formation of new oligomers but also the hydrolysis of previously formed oligomers during the “wet phase,” less-reversible cyclization of malic-acid-containing depsipeptides may have hindered further sequence diversification and chain elongation. Nevertheless, the chemical structure(s) of cyclo-dL and related products will be the subject of future study.

Depsipeptide chemical space maps enabled comparisons of relative abundance, composition, and hydrophobicity of the three types of linear depsipeptide systems examined. Relative abundances were found to vary across several orders of magnitude (Fig. 2A–C; log relative abundance corresponds to sphere size). For all three systems, species with primarily ester linkages, mixtures of ester and amide linkages, and exclusively amide linkages were observed (Fig. 2A–C; amide bond content corresponds to sphere color). It should be noted that exclusively amide-linked oligomers are technically not depsipeptides; for clarity purposes, however, these have been included under the depsipeptide umbrella. Depsipeptide hydrophobicity, evaluated via UPLC R<sub>t</sub>, varied as well. Leucine-containing
species typically eluted later in the chromatographic separations due to the more favorable association of its nonpolar side chain with the C18 stationary phase. IM measurements revealed an interesting linear correlation between neutral depsipeptide arrival times and oligomer length (SI Appendix, Fig. S6), indicating that the oligomers detected were likely unstructured. Deviation from such ion mobility trends could be useful in future searches of proteopeptide chemical space to find species with significantly different structures from the population mean, with such characteristics potentially indicating the emergence of secondary structure (40, 41).

Two-dimensional chemical space maps (Rt and dt) were also generated for one, two, and four environmental cycles to gain insight into depsipeptide evolution (Fig. 2 D-F). It was observed that, with increased number of cycles, a number of new glycolic-acid-containing depsipeptides were formed (yellow areas). For lactic-acid-containing species, new depsipeptides were formed as well, although within similar regions of Rt and dt space. As mentioned earlier, the length of malic-acid-containing species did not evolve appreciably, which is attributed to the formation of cyclic species. Overall, linear depsipeptides typically ranged from 2 to 8 residues in length, which is comparable to previous observations (12). Oligomers up to 14 residues in length have been previously observed (12), but such length typically requires a larger number of environmental cycles and/or a replenishing of monomers.

### D depsipeptide Sequence Isomers

The number of theoretically possible sequences increases greatly with oligomer length (Table 1). Therefore, we chose a multidimensional separation method with a peak capacity, on the order of $10^4$–$10^6$ (42–44). This method led to the detection of hundreds of sequences of glycolic-and lactic-acid-containing depsipeptides (Table 1), even separating sequence isomers on certain occasions (SI Appendix, Figs. S7 and S8). By nature, sequence isomers have identical precursor ion m/z values which are indistinguishable by the quadrupole mass filter used to isolate species before fragmentation. Despite our method’s peak capacity, a significant number of sequence isomers still overlapped in both UPLC and IM dimensions, resulting in coelution and MS/MS coselection. In the majority of these coselection cases, between two and five individual sequence isomers were found by manual interpretation of MS/MS fragmentation data. For longer depsipeptides (5–8 residues in length), up to 12 potential sequence isomers were found in selected cases. We also observed examples of longer depsipeptides fragmenting at internal ester linkages in the ion source region of the mass spectrometer, even under relatively soft electrospray ionization conditions (SI Appendix, Fig. S9). This is why a number of short sequences were detected multiple times in the same dataset (Table 1).

The relative abundances of certain species were found to change in a systematic fashion during the cycling process. The most commonly observed trend was the accumulation of new sequences with cycling. Sequences composed of >50% amino acid typically increased in relative abundance over time (selected examples in Fig. 3A), suggesting the rate of formation by ester condensation and subsequent ester–amide exchange outweighed the rates of hydrolysis and elongation into different species. The opposite trend was observed for ester-rich sequences of short-to-moderate length, as well as very short sequences (selected examples in Fig. 3B). Interestingly, we also observed species whose relative abundances were rather consistent over cycling (selected examples in Fig. 3C). We hypothesize such sequences, over this short time period, formed and hydrolyzed at relatively similar rates. Such species may also be hydrolysis products of longer, ester-rich sequences.

![Fig. 3.](https://example.com/fig3.png)

**Fig. 3.** Selected examples of changes in depsipeptide relative abundance after environmental cycling. (A) Coeluting sequence isomers dLa and dAL and coeluting sequence isomers gLGGg, gLGgG, and gGGLg increased in relative abundance with cycling. Enriched sequences contain primarily amino-acid residues with amide linkages. (B) Coeluting sequence isomers aAL and aLa and aA decreased in relative abundance with cycling. Sequences such as these are either short in length—likely involved in forming other species—or contain ester linkages. (C) Coeluting sequence isomers aaGGa and aGaGa and coeluting sequence isomers gAGg, gGAg, and gGgA did not change significantly between 1, 2, and 4 cycles. Sequences such as these may be intermediates in the formation of amide-rich sequences, but also may be hydrolysis products of longer, more ester-rich sequences.

### Comparing Theoretical and Experimental D depsipeptide Sequences

Homology diagrams (45) were generated for all theoretically possible combinations and experimental sequences of tetrameric species to examine sequence space as a function of the number of environmental cycles (Fig. 4). In these diagrams, the height of a letter corresponds to residue conservation. Taller letters indicate low sequence diversity (that is, low entropy) at a given position, whereas shorter letters indicate high sequence diversity (high entropy) at a position. As a visual comparison tool, the theoretical homology diagram shown in Fig. 4A is what would be obtained if all possible tetramers were produced in equal yields and differences in monomer reactivity were ignored. In this diagram, the likelihood of forming any backbone linkage, either...
Fig. 4. Comparison of theoretical and experimentally observed tetramer sequences using sequence homology diagrams. Larger letters indicate a residue is conserved and does not change appreciably with cycling. As letters become smaller, residues in the specific position become more random in nature and sequences more diverse. (A) Theoretical sequence distribution of a purely random mixture of depsipeptides containing a generic hydroxy acid (x) and glycine (G), alanine (A), and leucine (L) residues. Residues 2–4 would contain equal frequencies of x, G, A, and L monomers, assuming equal monomer reactivity. (B) Theoretical sequence distribution of a random mixture of x, G, A, and L depsipeptides in which all residue linkages are amide bonds, and the N-terminal residue is a hydroxy acid, regardless of sequence composition. Residues 2–4 would contain equal frequencies of G, A, and L monomers. (C) Sequence homology diagrams for the experimentally determined sequence distributions of g+G+A+L, a+G+A+L, and d+G+A+L tetramers after 1, 2, and 4 cycles. Both g+G+A+L and a+G+A+L tetramer sequences increase in their diversity with cycling. In contrast, d+G+A+L tetramers experience minimal change in sequence diversity with cycling. Malic acid-depsipeptides are enriched in amino acids, particularly leucine, at position 2.

The homology diagram shown in Fig. 4B is what would result if all ester linkages had been hydrolyzed and only amide linkages remained. In this scenario, all depsipeptides would still be initiated by a hydroxy acid, but would contain only amino acids of a random nature at positions 2–4. Although theoretical diagrams in Fig. 4A and B assume equal reaction rates between various amino acids for purposes of simplicity, recent kinetic studies by our team do suggest experimental differences in monomer reactivity (46).

Experimental sequence data for g+G+A+L, a+G+A+L, and d+G+A+L tetramers across one, two, and four environmental cycles were compiled into homology diagrams (Fig. 4C), and revealed the “N-terminal” residue to be a hydroxy acid for essentially all species, as expected from the proposed ester–amide exchange mechanism. Subsequent positions along the chain contained both amino acid and hydroxy acid residues, with varying degrees of randomness depending on cycle and hydroxy acid.

After one dry–wet cycle, glycolic-acid- and lactic-acid-containing tetramers contained little leucine. However, with additional cycling their sequence diversity increased, leading to approximately even mixtures of hydroxy acid, glycine, alanine, and leucine at positions 2–4. This suggests the reaction rate between leucine and glycolic or lactic acid may have been slower than that with glycine and alanine. Internal sequence stretches for g+G+A+L and a+G+A+L tetramers demonstrated high sequence diversity after only 4 d of reaction. The observation of more diverse tetramers as the system evolves is consistent with the hypothesis that a vast array of sequences may have been present on the prebiotic earth before the initiation of template-directed peptide synthesis (47, 48). When considering a prebiotic timescale, it seems reasonable to conclude that proto-peptides of short to moderate length would have had sufficient time to diversify into an innumerable set of sequences.

In contrast with g+G+A+L and a+G+A+L tetramers, the sequence diversity of d+G+A+L tetramers did not change significantly with cycling. Malic-acid depsipeptides contained more leucine than glycine or alanine after one cycle, and very few malic-acid depsipeptides were found to contain solely glycine and/or alanine without leucine (Fig. 4C). Position 2 of these depsipeptides appears similar to Fig. 4B in which few, if any, hydroxy acids would be present. This difference is attributed to the presence of nonlinear species in d+G+A+L mixtures which limit hydrolysis and elongation. In particular, the presence of only amino acids at position 2 is consistent with the formation of...
cyclo-DL species containing one ester and one amide bond. The chemical structure(s) of malic-acid-containing depsipeptides of a cyclic and/or branched nature will be the subject of future study.

Implications for Origins-of-Life Research
In this work, mixtures of depsipeptides were sequenced from three model prebiotic reaction mixtures. Hundreds of depsipeptides were formed from a total of six monomers over 4 d of dry–wet cycling, including a large number of sequence isomers. Tandem MS allowed for sequencing of species that coeluted in UPLC and IM separation space. Sequence homology diagrams demonstrated a high sequence diversity of depsipeptides containing glycolic acid or lactic acid, two of the most abundant hydroxy acids in model prebiotic reactions and meteorites (13, 14, 49). The diversity of these sequences as well as their peptide character increased with the number of cycles.

Prior studies on proto-peptide sequence diversity have taken either a purely computational approach (50) or a top-down approach in which libraries were made through existing biochemical means (51, 52). Such studies are informative, but could still benefit from complementary bottom-up studies to investigate the plausibility of the libraries themselves. This study provides direct experimental evidence in support of de novo proto-peptide libraries, and suggests the following concerning their composition: (i) overall sequence diversity would have been high; (ii) such diversity would have increased with time; and (iii) proto-peptide sequences capable of branching and/or cyclizing would have been less diverse than linear sequences due to competing reaction pathways.

For all systems studied here, it is important to note the number of confirmed sequences was lower than the number of theoretically possible sequences (Table 1). We attribute this to specific challenges associated with studying vast proto-peptide libraries using a bottom-up, de novo approach. An ever-present challenge of using MS-based approaches is that peptides must be detected with sufficient signal intensity to be sequenced. Differences in peptide ionization efficiency play a role in overrepresenting species with efficient signal intensity to be sequenced. Differences in peptide MS/MS can influence the number and confidence of sequence assignments could decrease particularly for proto-peptide mixtures like the ones examined here which contain a number of sequence isomers. It is possible that using DIA methods, or methods which combine aspects of both DDA and DIA (56), could enhance sequencing capabilities.

An additional challenge is related to the chemistry used to make de novo proto-peptides: model prebiotic reactions which occur over short timescales may bias our understanding of plausible sequences toward species which are most kinetically favored. While it is obviously impossible to perform experiments on a prebiotic timescale, it is important to consider reaction time, or in this instance the number of evaporative cycles, as a key variable in generating proto-peptide libraries. Our data demonstrate that glycine and alanine residues incorporated into glycolic-acid- and lactic-acid-based depsipeptides earlier than leucine (Fig. 4C, left two columns). These differences in reactivity should be further studied in detail to gain a clearer picture of the proto-peptide inventory. Also, both the consumption of free monomer(s) and the evaporation of hydroxy acids upon heating have been shown to affect the observed product distributions (46). Better feed approaches to replenish monomers in ester–amide exchange reaction cycles, and more diverse monomer sets, would thus be expected to produce even higher sequence diversity.

It has been recently argued that origins-of-life studies concerning proto-peptides are often too narrowly focused on biologically coded amino acids, and should consider the vast diversity of nonproteinogenic monomers that would have also been present in the prebiotic milieu (57, 58). Our use of hydroxy acids to form depsipeptides demonstrates the utility of this approach. MS-based approaches may be used in the future to explore the vast chemical space of plausible proto-peptides containing not only hydroxy acids and proteinogenic amino acids but also nonproteinogenic amino acids, or mixtures of the three. As long as a nonproteinogenic monomer has a unique m/z among the other monomers in the reaction, it should be readily identifiable within peptide sequences. Exceptions would be structural isomers of nonproteinogenic amino acids (e.g., β-alanine and norvaline) which are isobaric with their proteinogenic forms (α-alanine and valine, respectively); however, it is possible IM could resolve these species (59). As broader sequence space is explored, it is possible that catalytic peptides—and perhaps even autocatalytic species—may arise which directly increase proto-peptide length and function (60).

In addition to using a broader set of monomer residues, other molecules and ligands such as proto-nucleic acids, proto-lipids, proto-glycans, and inorganic species could be introduced as well. Such molecules would have certainly influenced the solubility, formation, and hydrolysis of proto-peptides, and may have coevolved together with proto-peptides in primordial environments (61). Experimental observation of proto-peptide function such as catalysis or self-replication within a more realistic model of the prebiotic soup will provide a more robust picture of chemical evolution in an origins-of-life context, highlighting systems-level interactions within highly complex chemical networks (62).

ACKNOWLEDGMENTS. We thank Prof. Matthew Torres and Dr. Nagender Panyala (Georgia Institute of Technology) for advice on sequence logos and proteomics-style data analysis, as well as Mark Bennett (Nonlinear Dynamics) for assistance with commercial MS software. This work was supported by the NSF and the NASA Astrobiology Program, under the NSF Center for Chemical Evolution (Award CHE-1504217).


