Prebiotic model of peptide formation based on molecular weight and thermal decomposition temperature

C. Polanco[1], **T. Buhse**[2], **J.L. Samaniego Mendoza**[1], **A. Morales Reyes**[3], **M. Arias Estrada**[3] and **V.N. Uversky**[4,5]

[1] Department of Mathematics, Faculty of Sciences, Universidad Nacional Autonoma de México, México City 04510, México
[2] Centro de Investigaciones Químicas, Universidad Autónoma del Estado de Morelos, Cuernavaca Morelos 62209, México
[3] Department of Computer Science, Instituto Nacional de Astrofísica, Óptica y Electrónica. Tonanzintla Puebla, 72840, México
[4] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33647, USA
[5] Protein Research Group, Institute for Biological Instrumentation of the Russian Academy of Sciences, 142290, Pushchino, Moscow region, Russia

Keywords: Abiogenesis; polymerisation; origin of life; molecular weight; thermal decomposition

In this work, we present a stochastic computational model that recreates the formation of prebiotic peptides from a set of 20 proteinogenic amino acids, taking into account their molecular weight, polarity, and thermal decomposition. The model simulates the recombination, fragmentation, and self-replication of the peptides. The results show great similarity between the polar profile of these peptides and those obtained by this team in a simulation of small peptide generation based on the amino acid abundances observed in the classical Miller–Urey experiment of prebiotic amino acid formation and the most conserved genes found in three microorganisms, *Methanococcus jannaschii*, *Escherichia coli* and *Saccharomyces cereviasiae*. This leads to the assumption that in a natural biopolymerisation process, the molecular weight and the relative abundance of the amino acids might serve as important factors driving the formation of primordial peptides. The molecular weight, which can be used as a proxy for the degree of structural complexity, is a factor that can influence the synthesis and abundance of amino acids. In order to determine this polar profile, we used the bioinformatics Polarity Index Method©, which has also been used for the characterization of different groups of peptides and proteins.

1. Introduction

The study of structural and functional protein profiles has been, and still is, a fundamental subject in the research dedicated to finding structure–function correlations, manufacturing new pharmaceutical drugs and understanding the mechanisms of disease [1, 2]. It also helps deepen the understanding of the proteinogenic polymerisation mechanism utilized by Nature in the prebiotic world.

This paper introduces an abiogenetic model that re-creates the biopolymerisation process leading to the formation of the first peptides (four billion years ago) from a set of 20 proteinogenic amino acids, without considering the evidence obtained in the Miller–Urey experiments on amino acid formation [3], Fox and Harada's experiment on the formation of proteinoids [4], Rode's experiment on salt-induced dipeptide formation [5], or analysis of fossils from the Ordovician period (430,000,000 years ago) [6]. This stochastic computational model only ponders a distinctive variable of the amino acids, their molecular weight [7], and re-creates the dynamics from the polarity difference [8], owing to the variability of amino acid reactive groups and their thermal decomposition [9]. These variables were used to modify an evolutionary computational platform, developed and published by our team, that mimics the abiogenetic formation of peptides. The peptides generated in our current study were compared with those obtained in a model based on the results of the Miller–Urey experiment [10], and with a set of proteins encoded by highly conserved genes [11] from three microorganisms: *Methanococcus jannaschii*, *Escherichia coli* and *Saccharomyces cereviasiae*. The metric used to compare these sets of peptides was the polar profile, an approach that was proven to be efficient in the discrimination of different groups of peptides and proteins and that was used for the development of a unique bioinformatics tool, Polarity Index Method [10]. The results of this analysis show a high correlation between the two sets of peptides studied, namely, peptides generated in this study, and peptides based on the results of the Miller–Urey experiment. Our observations allow one to assume that four billion years ago (in the prebiotic world), the incidence of amino acids present in the peptides could be significantly influenced by their molecular weights (i.e. their degree of structural complexity).

2. Materials and Methods

2.1 The Model

The stochastic-computational model produced peptides of 25 amino acids in length from a set of 20 proteinogenic amino acids (Table 1) and recreated the polymerisation of peptides using two factors: recombination and self-

replication. Therefore, in this model, the polypeptide formation depended on the molecular weight, thermal decomposition, and polarity of the amino acids. The polymerisation process randomly utilises amino acids, which are added to the peptide according to three parameters: (i) the charge of the amino acid that will be added and the charge of the terminal amino acid in the peptide (see 2.1.1 Polarity section), (ii) the molecular weight of the amino acid that will be added (see 2.1.2 Molecular weight section) or (iii) the thermal decomposition temperature (see 2.1.5 Thermal decomposition temperature section). This utilisation of amino acids for polymerisation, whose weighting (i–iii) determines their addition to the forming peptide, is further affected by the modelling of the proteolysis process, which starts randomly in two ways: (a) when the peptide is cut and one part continues the polymerisation process, whilst the other remains available to merge with another peptide (see 2.1.3 Recombination section) and (b) when a new peptide emerges from the copy of the part forming the peptide (seed peptide) (see 2.1.4 Self-replication section). Both processes (a and b) reproduce the evolutionary aspect of polymerisation as a combination of heritage and chance. The simulation of all these processes and variables requires a short-memory stochastic scheme so the modelling complexity does not increase when the number of variables increases. Our model is a stochastic process and although it has a strong similarity with a Hidden Markov model, it was not computationally simulated as such. This similarity is discussed in detail in the 2.1.6 Stochastic profile section. The execution of this computational model (see 2.1 Model section) was intended to alternatively enable and disable three parameters (polarity, thermal decomposition temperature, and molecular weight) in order to determine the dependency and significance of each parameter. The peptides obtained with this simulation are evaluated with the computational method called Polarity Index Method (see 2.3 Polarity Index Method section). This method calculates the relative frequency of each of the 16 polar interactions [12, Table 2] resulting from the linear sequence of each peptide when reading it in pairs from left to right, counting the polar incidences that result when each pair of amino acids is read, moving one figure at a time. The relative frequencies thus obtained are expressed as histograms (see Figures 1–3).

Table 1 Amino acids used in the simulations.

Amino acid	**Symbol**	**Molecular weight** [a]	**Thermal decomposition temperature (°C)** [b]	**Polarity** [c]	**Amino acid**	**Molecular weight (g/mol)** [d] **(ascending order)**	**Number of Times** [e]
Alanine	A	89.1	295	NP	G	75.1	2
Arginine	R	174.2	246	P^+	A	89.1	3
Asparagine	N	132.1	238	NP	S	105.1	23
Aspartic acid	D	133.1	270	P^-	P	115.1	41
Cysteine	C	121.2	231	P^-	V	117.1	71
Glutamic acid	E	147.1	210	P^-	T	119.1	102
Glutamine	Q	146.1	196	NP	C	121.2	**130**
Glycine	G	75.1	259	N	I	131.2	158
Histidine	H	155.1	288	P^+	L	131.2	166
Isoleucine	I	131.2	284	NP	N	132.1	**284**
Leucine	L	131.2	294	NP	D	133.1	403
Lysine	K	146.2	233	P^+	Q	146.1	**464**
Methionine	M	149.2	289	NP	K	146.2	526
Phenylalanine	F	165.2	276	NP	E	147.1	987
Proline	P	115.1	231	NP	M	149.2	**1349**
Serine	S	105.1	232	N	H	155.1	**1711**
Threonine	T	119.1	259	N	F	165.2	**2073**
Tryptophan	W	204.2	293	N	R	174.2	**2435**
Tyrosine	Y	181.2	316	N	Y	181.2	**2797**
Valine	V	117.1	315	NP	W	204.2	**3159**

[a] Molecular weight of each amino acid [7]. [b] Temperature at which the substance chemically decomposes [9]. [c] Classification of amino acids by their charge: acidic hydrophilic (P^-), basic hydrophilic (P^+), neutral (N), and non-polar (NP) [8]. [d]Molecular weight ofeach amino acid [7]. [e]Number of times according to the Miller-Urey distribution [12, Table 2], is the loop index that the model uses, to decide whether or not to add an amino acid to the protein under construction. Note: the bold numbers were estimated (see 2.1.2 Molecular weight section).

The panels on the left and on the right in [12, Table 2] represent two equivalent quantitative representations of the polar interactions between the charges. On the left panel are given the empirically polar interactions previously calculated by us and the right panel represents their inverse percentiles. The model here introduced uses the values of the right panel.

2.1 Polymerisation

2.1.1 Polarity

Our computational simulation recreates the formation of a peptide starting with the random formation of a dipeptide, then the model randomly produces two amino acids. The peptide bond between these random amino acids is formed when the NH_2 group of the second amino acid combines with the -COOH group of the dipeptide, depending on the polar interaction between the newly generated amino acid and the amino acid located at the end of the first dipeptide. The model encouraged some interactions more than others according to the corresponding weighting [12, Table 2]. The most encouraged polar interactions were $[P^+,P^-] = [P^-,P^+]$ and the least encouraged were $[P^-,P^-] = [P^+,P^+]$.

2.1.2 Molecular weight

After an amino acid was randomly generated and before comparing its charge, the amino acid was considered according to its molecular weight. Thus, our computational model favoured the amino acid formation based on its molecular weight. In this way, amino acids with lower molecular weight (Table 2) had a greater opportunity to prevail, and eventually join the peptide chain, than those with higher molecular weight. The "number of times" distribution is empirical and pretended to follow roughly the ratio of the Miller-Urey model (Figure 1).

2.1.3 Recombination

This process recreated proteolysis [12], inducing cuts of the peptide according to the following criterion (1), where e = 2.7183 and L is the peptide length at any particular moment.

$$C\ (L) = 1/e^L \qquad (1)$$

Note: The peptide cutting is due to this function. The selection of the peptide to be divided follows a random process. As a result, the original peptide was divided into two sub-peptides, one of which took the place of the original peptide and continued the polymerisation process to the end, whilst the other segment was added to a pool of peptides where it might (or might not) join another peptide that was forming in an inverse process.

2.1.4 Self-replication

Our computational model was designed to simulate random peptide self-replication. To achieve this aspect, the model independently generates peptide "parents" and "children" in sub programs with thier own and independent characteristics. Some sub programs started from a randomly generated dipeptide and other times it took a copy of any of the segments of the forming peptide and “sent” it as a “seed” to the new process. From a computational perspective, the self-replication process started with one process that ran on a computer, followed by two or more processes running simultaneously, on the same computer. Self-replication gave the model the possibility to observe the generational change of the peptide composition in a range of 15 generations.

2.1.5 Thermal decomposition temperature

The model decreased the molecular weight of each amino acid from the second generation by a factor of 0.0001 × thermal decomposition temperature (Table 1). The determination of this factor (0.0001 × thermal decomposition temperature) is empirical.

2.1.6 Stochastic profile

Our computational simulation recreates a stochastic process that resembles a Hidden Markov Model (HMM), where the molecular weight, thermal decomposition, and polarity are the visible variables. The polar profile of the peptide (generated by our computational simulation) is the hidden variable, and peptide generation is assumed to be a Markov process [15-20]. According to Ching [15] a HMM features: (i) states (polar profile of peptide), (ii) visible variables (molecular weight, thermal decomposition and polarity), (iii) start probability (initial polar profile), (iv) transition_probability (proposed molecular weight, thermal decomposition and polarity), and (v) emission_probability (polar profile of the generated peptide). It is important to note that our computational model is not a HMM, but a stochastic model, because it generates random proteins, from a set of physicochemical variables.

2.1.7 Computational specifications

For the processing of the information we used a computational platform consisting of HP Workstation z210 — CMT — 4 x Intel Xeon E3-1270/3.4 GHz (Quad-Core) — RAM 8 GB — SSD 1 x 160 GB — DVD SuperMulti — Quadro 2000 — Gigabit LAN, Linux Fedora 27, 64-bits. Cache Memory 8 MB. Cache Per Processor 8 MB. RAM 8 GB. The computational implementation of our model required building a program in Fortran 77 with Linux scripts, with external calls to the operating system. It had to be capable of self-replication with different random parameters to start and divide the protein and simulate the evolutionary process of reproduction from father to offspring. In order to achieve this, we used a random number generator of the linear congruential type, with six different seeds in each program. With the internal processes of the sub programs that randomly decided the self-reproduction, we faced the problem of data overflow by the number of self-generated processes. Therefore, we added a supervisor program that randomly closed some of the sub programs. Reproducing up to 15 generations took four days, a computer platform with fewer processors overflows and eventually, the process needs to be re-started. The final number of generated proteins was 333377 and it was observed that self-replication had a distribution (from generation 0 to 15): 119_0, 888_1, 3547_2, 8146_3, 12601_4, 14642_5, 12014_6, 6467_7, 11499_8, 19929_9, 32559_{10}, 40173_{11}, 49080_{12}, 51653_{13}, 40587_{14}, 29473_{15}.

2.2 Miller–Urey experiment

The computational simulation of the Miller–Urey experiment was previously reported by this group [12] using a set of 21 amino acids and weighting two variables: the polarity (Table 1) and the abundance of the amino acids (Table 1), and we modified both variables and identified that the abundance of the amino acids had a preponderant role in the polymerisation of the first peptides (four billion years ago).

2.3 Polarity index method

In order to identify differences, we used the bioinformatics Polarity Index Method [10], which produced a graph called a polar profile. This method involved the following steps: (i) The "n" peptides (each one of length 25 amino acids) were assembled in a single 25n length amino acid peptide. (ii) The amino acids were replaced by their polar numerical equivalence, i.e. $[P^+] = 1$, $[P^-] = 2$, $[N] = 3$, and $[NP] = 4$. (iii) An incidence matrix was generated (4 rows × 4 columns), recording the polar incidences reading from left to right, each pair of amino acids one position at a time. (iv) The elements of the incidence matrix were normalised to one. (v) The incidence matrix was linearised in a vector of 1 × 16, allocating the first line of the matrix to the first four positions of the vector, the second line to the second four positions, and so on. The vector represented the relative frequencies of the 16 polar interactions. (vi) The vector was plotted; the Y axis represented the relative frequencies (%) and the X-axis the 16 polar interactions. This graph was defined as the polar profile (Figure 1).

Example. As mentioned in this section, the procedure starts with the selection of an amino acid set and its numerical classification into four polar groups. For the peptide SYVDHLMCDVE, its numerical equivalence would be 33421443242.

This new numerical sequence is read from left to right, in pairs, one amino acid at a time, and each pair found is recorded in an incidence matrix M[i,j], where i is the row, and j is the column.

1) Initialise M[i,j] matrix i.e. M[i,j] = 0, for all I,j = 1, 4. The numerical sequence is read from left to right. The first pair found is "33", which is equivalent to row i = 3 and column j =3 in M[3,3] matrix so "1" value is added to M[3,3]. Then one position is moved to the right and pair "34" is found so in row i = 3, column j =4 so "1" value is added to M[3,4] matrix. The same procedure continues to the end of the sequence, when all incidences are registered in the M[i,j] matrix.

2) Step 1 is repeated for all peptides in the "training protein group", and all incidences are added to the matrix **M[**i,j**]**. At the end, matrix **M[**i,j**]** is normalised. Since the Polarity Index Method is a computational supervised method, it needs to be trained. To reach this objective it is necessary to have a representative peptide set (or training protein group) with the characteristic to find. Please note that the initialisation of the **M[**i,j**]** matrix is performed only at the beginning of this procedure and not with each peptide, i.e. this incidence registration is cumulative. The next step is to normalise the **M[**i,j**]** with all the amino acids found in the "training protein group".

3) Steps 1–2 are repeated for each peptide studied and its incidences are registered in the **B[**i,j**]** matrix. Finally, each **B[**i,j**]** matrix is weighted, i.e. matrix **B[**i,j**]** + matrix **M[**i,j**]** = matrix (**B[**i,j**]** + **M[**i,j**]**).

4) This last paragraph means that if four peptides are evaluated, steps 1–3 are repeated for each peptide, resulting in four $(\mathbf{B[}i,j\mathbf{]}_k + \mathbf{M[}i,j\mathbf{]})_k$ matrices at the end of this procedure. The Polarity Index Method compares each position of each element (i,j) in both matrices. The number of coincidences or matches is interpreted by the method as the "level of similarity" and is expressed as a percentage of the number of hits.

5) The final step is to compare each $(\mathbf{B[}i,j\mathbf{]}_k + \mathbf{M[}i,j\mathbf{]})_k$ matrix with the **A[**i,j**]** matrix. Suppose the highest relative frequency in both matrices is (i,j) = (1,2); this means the number of hits up to that moment is 1 out of 16, i.e. (1/16 = 6.25%). If the second highest relative frequency in both matrices is located in the same position, the number of hits will be 2 out of 16, i.e. (2/16 = 12.50%), the procedure continues until the lowest relative frequency is determined.

2.4 Conserved microorganisms

The most evolutionary conserved genes from these three microorganisms: *M. jannaschii, E. coli* and *S. cereviasiae,* which are considered as the so-called common ancestors of approximately 2.5 billion years ago [11], were taken and their polar profiles were calculated.

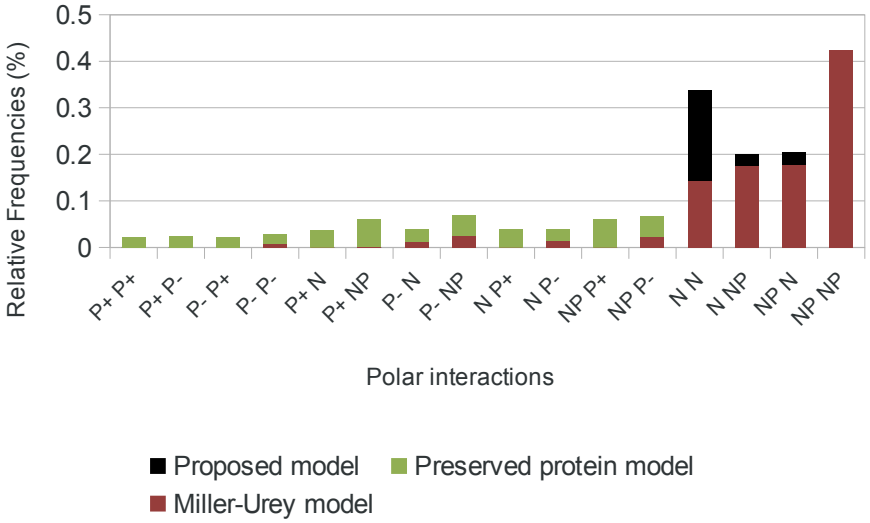

Fig. 1 Polar profiles of the model, the Miller-Urey experiment and the most preserved microorganisms. Histogram. The X-axis represents the 16 polar interactions [10].

2.5 Test files

The tests of the peptides generated were applied to determine the following:

(a) The correlation (or lack of it) between variable molecular weight and thermal decomposition temperature in our computational approach.
(b) The differences of the peptides produced by our computational simulation for each generation (see 2.1.4 Self-replication section).
(c) The polar profile of peptides generated by our computational simulation.
(d) The differences in the polar profiles (see 2.3 Polarity Index Method section) from our computational simulation here described and our computational simulation of the Miller–Urey experiment [16].
(e) The polar profile differences between the set of peptides generated by our computational re-creation and the peptides from the most preserved genes found in three microorganisms: *M. jannaschii, E. coli* and *S. cereviasiae*.
(f) The polar profile differences between the set of peptides generated by our computational simulation of the Miller–Urey experiment using the 21 amino acids {G, A, S, P, V, T, I, L, D, K, E, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9} [12] and the same computational recreation but only using the proteinogenic amino acids {G, A, S, P, V, T, I, L, D, K, and E} (Table 2).

Discrete data are customarily interpreted with bar graphs but we have interpreted them as continuous data here in order to appreciate the general tendency among both distributions.

3. Results

When the amino acids used in the model are sorted in ascending order by their molecular weight and thermal decomposition temperature, there is a match for 12 of the 20 amino acids (Table 2).

The polar profile of the model and the polar profile of the Miller–Urey experiment [12] (Figure 1) show a correlation in 14 of the 16 polar interactions, except for the inflection points of [P^-, N] and [N, P^+].

The polar profile from the Miller–Urey experiment and the polar profile of the peptides of the most conserved genes [12] show a correlation in 15 of the 16 polar interactions, except for the polar interaction [P^-, NP], making translation of the graphs obvious (Figure 3).

Table 2 Molecular weight and abundance.

Molecular weight and Miller-Urey abundance					**Molecular weight and Thermal decomposition**				
Molecular weight (g/mol) [(a)] (**ascending** order)	Amino acids ←	Matches ± 2 positions [(b)]	Proteinogenic amino acids	Miller-Urey experiment [(c)] (**descending** order)	Molecular weight (g/mol) [(a)] (**ascending** order)	Amino acids ←	Matches ± 2 positions [(b)]	Amino acids →	Thermal decomposition temperature [(c)] (**ascending** order)
75.1	G	▪	A	790	75.1	G		Q	196
89.1	A	▪	G	440	89.1	A		E	210
105.1	S		D	34	105.1	S	▪	C	231
115.1	P		V	19.5	115.1	P	▪	P	231
117.1	V	▪	L	11.3	117.1	V		S	232
119.1	T		E	7.7	119.1	T	▪	K	233
131.2	I	▪	S	5	121.2	C	▪	N	238
131.2	L	▪	I	4.8	131.2	I		R	246
133.1	D		P	1.5	131.2	L		G	259
146.2	K	▪	K	1.2	132.1	N	▪	T	259
147.1	E		T	0.8	133.1	D	▪	D	270
Molecular weight (g/mol) [(a)] (**ascending** order)	Amino acids ←	Matches ± 2 positions [(b)]	Proteinogenic amino acids	Miller-Urey experiment [(c)] (**descending** order)	146.1	Q		F	276
75.1	G	▪	A	790	146.2	K	▪	I	284
89.1	A	▪	G	440	147.1	E		H	288
105.1	S		D	34	149.2	M	▪	M	289
115.1	P		V	19.5	155.1	H	▪	W	293
117.1	V	▪	L	11.3	165.2	F	▪	L	294
119.1	T		E	7.7	174.2	R		A	295
					181.2	Y	▪	V	315
					204.2	W	▪	Y	31
[a]Molecular weight of each amino acid [7]. [b]Distance between amino acids, (abs) (amino acid-a, amino acid-b) ≤ 2 positions. (c) Abundance of amino acids according to the Miller-Urey experiment [12, Table 1].					[a] Molecular weight of each amino acid [7]. [b]Distance between amino acids, (amino acid-a, amino acid-b) ≤ 2 positions. [c]Temperature at which the substance chemically decomposes [9].				

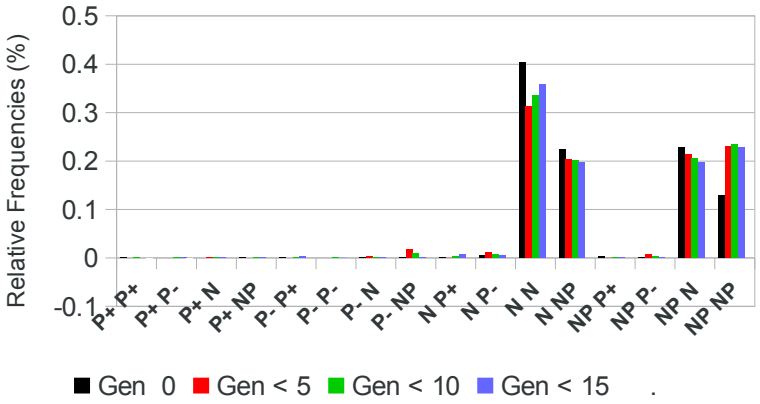

Fig. 2 Polar profile histogram of the model through 15 generations. The X-axis represents the 16 polar interactions [10].

The polarity profiles of the 15 generations produced by the model (Figure 2) show a close coincidence of the 16 polar interactions, except for the polar interaction [N, P^+]. The thermal decomposition temperature does not seem to affect the distribution of the polar profiles.

When assorting the amino acids in ascending order by their molecular weight and abundance according to the last simulation of the Miller-Urey experiment [10, Table 1], there is a correlation in six of the 11 amino acids (Table 2; except S, P, T, D, and E).

The zero generation of the polar profiles in the Miller-Urey computational model, with and without proteinogenic amino acids, showed no difference.

4. Discussion

Computational models aimed at reproducing experiments related to the origin of life usually begin by determining the number and magnitude of the variables involved. The difference in magnitude of these variables allows the verification of the assumptions made. Although this model does not have an experimental process, it can essentially be described as such because it takes a set of 20 amino acids (Table 1) and simulates the polymerisation process of proteins. It considers two primary processes—recombination and self-replication—and combines them with three properties inherent to the amino acids: molecular weight, thermal decomposition temperature, and polarity. This kind of stochastic process that resembles a Markov stochastic scheme, where none of the variables in the computational model can have a fixed value to observe the behaviour of the others, so it requires the participation of all variables simultaneously. The complexity in the biological simulation comes from the number of variables included that cannot have a determined value. In this sense, our model benefits from its resemblance to a Hidden Markov model, where the Markovian stochastic models are more suitable than the differential-deterministic models. The fundamental difference in this modelling was the introduction of the molecular weight and thermal decomposition temperature of each amino acid, instead of a fixed value for the abundance of each amino acid based on fossil evidence [6], meteorites [3], and salt-induced peptide formation [5]. This change in our computational simulation was based on the conjecture that the abundance of the amino acids present in prebiotic proteins was, and it still is, the result of nature's "preference" to build amino acids with low molecular weights and thermal decomposition temperatures. Our results show this in 12 of the 20 amino acids where both factors are compared (Table 2) and in the graph (see Figure 2) where the polar profiles of the proteins obtained in a simulation of small peptide generation based on the amino acid abundances observed in the classical Miller–Urey experiment [12] of prebiotic amino acid formation and our model are compared, as well as in the comparison with the proteins of the most preserved genes found in the three microorganisms [11].

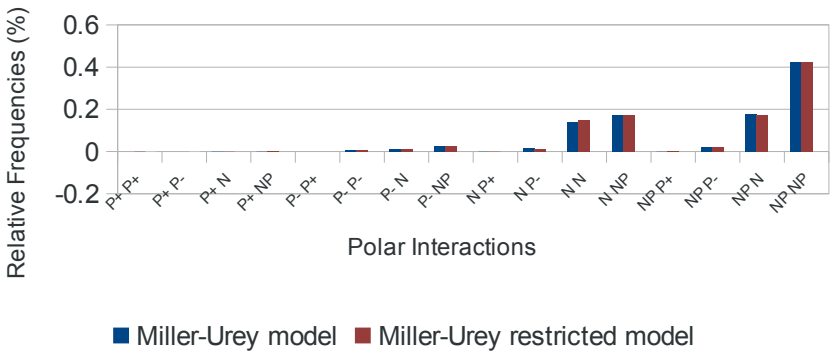

Fig. 3 Polar profile histogram of the Miller-Urey model restricted to amino acids {G,A,S,P,V,T,I,L,D,K,E}, and with 21 amino acids [12]. The X-axis represents the 16 polar interactions [10].

There is an almost complete correlation (14 to 16 polar interactions) between the polar profile of the Miller-Urey computational model using the 21 amino acids [12, Table 1], and the same model restricted to the proteinogenic amino acids {G, A, S, P, V, T, I, L, D, K, and E} (14 to 16 polar interactions over 50%: [P^+,P_+], [P^+,P^-], [P^-,P^+], [P^-,P^-] [P^+,N], [P^-,N], [P^-,NP], [N, P^+], [N, P^-], [NP, P^-], [N, N] [N,NP], [NP,N], and [NP, NP]; Figures 4.a, 4.b, and 4.c). This can be understood in the sense that the non-proteinogenic amino acids {0, 1, 2, 3, 4, 5, 6, 7, 8, and 9} have lower abundance, except amino acids 1, 8, and 9 [12, Table 1]; therefore, their participation in the modelling is much lower. It is important to mention that not all assumptions raised when affecting a variable are met since it is a stochastic process and all variables are randomly changing. A very different scenario occurs in deterministic models. It is important to note that our results are not conclusive. The amino acids E and D (Table 2) occupy extreme positions in our computational recreation of the Miller–Urey experiment as well as in our simulation described here. However, this work shows that there is a strong association between the molecular weight of the amino acids and the incidence observed in the synthetic peptides produced by our computational modelling, as well as in the more consolidated peptides found in the microorganisms used in this study. The Miller–Urey experiment [7] identified a number of amino acids by water phase, whilst our computational simulation identifies possible factors that could have some relation to the generation of those amino acids using the chemical decomposition caused by heat. We consider that the two approximations are not opposite. Our model aims to provide useful information related to the first polypeptides (four billion years ago), showing the possible factors related to the polymerisation process, using the metric of "polar profile" to identify the evolution of those first peptides. What is the probability that the similarities are not due to random coincidence? We have a graphical coincidence in 14 of 16 polar interactions (Figure 1), 14 / 16 = 87%. It is high but not definitive, even though this computational approach shows that the thermal decomposition temperature and molecular weight play important roles in the polymerisation process. For the sake of simplicity, we considered in our computational simulations only the biologically most relevant α-amino acids in the assumed prebiotic scenario. Keeping in mind that the Miller–Urey experiment yields a racemic mixture of amino acids, no distinction was made between L- and D-amino acids. However, the present simulation work is directed to include the effect of possible enantiomeric imbalances, which is in line with our recent publication devoted to the evolution of non-racemic peptides from a racemic pool of amino acids [21]. Our computational model refers to the dynamic formation of peptides from a pool of amino acids with individual concentrations estimated from the classical Miller-Urey experiment. Hence, the model considers that amino acids are already present in its initial stage. Nevertheless, the amino acids produced during a Miller–Urey type prebiotic scenario were additionally assumed to be subjected to thermal decomposition processes during evolutionary time scales that alter their individual concentrations. These thermal decomposition processes are plausible for the prebiotic Earth because of the presence of energy sources such as volcanic activity, lightening events or hot thermal springs. Hence, the initial amino acid concentrations estimated by Miller-Urey's experiments became a function of their thermal stabilities on the prebiotic Earth. This leads to normalised concentrations in the amino acid pool for the simultaneously ongoing peptide formation. How can these (and future) results progressively complement the cumbersome efforts of experimental chemistry to fathom the intrinsic complexity of life-like self-organization on a pristine early Earth? The production of these drugs partially comes from the identification of peptides and proteins in nature, but another alternative comes from the production of synthetic proteins, whose manufacture is to alter (adding or removing), amino acids from the linear sequence of the peptide or protein, or build it from scratch. In both cases all knowledge about the linear representation of proteins we consider to be a valuable contribution in the direction of understanding how nature constructs these essential functional units

4.1 Computational options: parallel architectures

Nowadays, it is possible to access off-the-shelf massive parallel processing platforms. Such technologies are able to deploy fine and coarse grained parallel schemes, depending on communication costs and algorithmic constraints. The self-replication algorithm used by the proposed computational model (see Subsection 2.1.4) requires to simultaneously launch a number of processes executing the same set of algorithmic steps; each process requires to read and write same data (tight bottlenecks emerge) and different data from/to memory implying also a vast amount of memory resources. It is common to use a Single Program Multiple Data (SPMD) technique to achieve parallelism either in fine or coarse grained processing platforms [22]. This technique allows a decentralized control to distribute the workload among processor elements that independently execute a set of instructions on locally stored data reading and writing results from/to higher memory levels in order to avoid bottlenecks and racing conditions [23]. Incorporating such a parallel scheme would allow for the proposed model to evolve a further number of generations and thus deepen observations throughout the evolution process.

4.2 Microprocessing: advantages and disadvantages

The algorithm presented here is based on a GA (genetic algorithm)-like formulation and therefore it is suited for parallelization. Increasing the performance will help to explore more iterations and arrive to more in-depth insights of the research. In order to do an efficient parallel implementation, the programmer needs to consider not only the code

efficiency but also the limitations of the parallel architecture and the memory access bottlenecks, as mentioned in the previous section. A popular solution is the use of CUDA (Compute Unified Device Architecture) [24], which is based on commercial off-the-shelf GPUs. The CUDA language is an abstraction of the parallelization that can be done in the GPU architecture, and is based on the C language, where the programmer is exposed to an abstraction of the architecture and local/global memories, but hides the actual threads scheduling and the level of parallelism which depends on the model of the GPU board. There are several guidelines and ways to reformulate an algorithm to achieve parallelism that is proposed in modern CUDA language courses [25, 26], and genetic algorithms is a class of algorithms that can benefit of CUDA programming and achieve high performance on a desktop computer. Early work in GA algorithm CUDA implementation [27] presents strategies to map the algorithm into the CUDA software model, with speedups of 1000 or more. In GA algorithms, the actual limiting factor is the complexity of the evaluation function, which will be executed in all the threads and limited by the power of the actual individual processing unit in the GPU. In our case this is not an issue since the calculations are relatively simple and short per thread. Other researchers have achieved efficient scheduling [28] of threads taking advantage of the latest GPUs architectures with minimum code modification. Further work on our research simulation will focus on using GPUs to increase the number of iterations with an accelerated and efficient reformulation of the code.

4.3 Bioinformatics and synthetic proteins

The design and construction of proteins from bioinformatics algorithms based on specific physico-chemical properties seem not to require the evolutionary knowledge of proteins but only the training of the protein set. However, the truth is that this discipline would greatly benefit from this knowledge by introducing the use of the proteins found in nature, whose interaction with other living organisms and proteome is already known. This would minimize the risk of constructing synthetic proteins without a protein profile similar to the proteins found in nature. The bioinformatics methods used for the design and construction of synthetic proteins should take into account the protein profiles already available before offering a protein that will solve the immediate problem but will have an unknown effect on the proteome.

5. Conclusions

The model presented in this study takes the inherent properties of the amino acids, independently of the circumstantial abundance and other experiments. The relevance of these results lies in the high similarity between the graphs of the polar profile of the model, the Miller–Urey experiment, the most conserved genes and our computational model based on thermal decomposition temperature. These results suggest that our model partially reproduces a likely prebiotic scenario (four billion years ago). We also observed that there are 12 of 16 matches (Table 2) when sorting amino acids in ascending order by molecular weight and thermal decomposition temperature (Table 2), suggesting that the thermal decomposition temperature is a factor that could decidedly influence the relative abundance of the amino acids since a lower thermal decomposition temperature is related to their greater relative abundance.

Supplementary Materials The programs used in this work can be requested from the corresponding author Carlos Polanco (polanco@unam.mx).

Author Contributions Theoretical conceptualization, design, and computational performance: CP. Data analysis: CP, TB, AMR, MAE, JLSM, and VNU. Results discussion: CP, TB, MAE, AMR, JLSM, and VNU.

Funding This research received no external funding.

Acknowledgments We thank Concepción Celis Juárez whose suggestions, and proof-reading have greatly improved the original manuscript, and also we acknowledge the Computer Science department at Institute for Nuclear Sciences at the Universidad Nacional Autónoma de México for support.

Conflicts of Interest The authors declare no conflict of interest.

References

[1] Ji HF, Kong DX, Shen L, Chen LL, Ma BG, Zhang HY. Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. Genome Biol 2007;8:R176.
[2] Julian D, Dorothy D. Origins and Evolution of Antibiotic Resistance. Microbiol Mol Biol Rev 2010;74:417–433. DOI: 10.1128/MMBR.00016-10.
[3] Miller SL. A Production of Amino Acids under Possible Primitive Earth Conditions. Science 1953;117:528-529.
[4] Fox SW, Harada K. Thermal copolymerization of amino acids to a product resembling protein. Science, New Series 1959; 128: 1214.

[5] Rode BM. Peptides and the origin of life. Peptides 1999;20:773-786.
[6] Abelson PH. Chemical events on the primitive earth. Proc Natl Acad Sci USA 1966;55:1365-1372.
[7] The index Merck. An Encyclopedia of chemicals, Drugs, and Biologicals 13 Ed. Merck and Co, Inc Ed. 2001.
[8] Pauling L. General Chemistry [Química General] 3rd edition W. H. Freeman & Company Publishers, 1955, pp. 227,621.
[9] Olafsson PG, Brynn AM. Evaluation of thermal decomposition temperatures of amino acids by differential enthalpic analysis. Mikrochim Acta 1970;5:871-878.
[10] Polanco C, Buhse T, Samaniego JL, Castañón González JA. Detection of selective antibacterial peptides by the Polarity Profile method. Acta Biochim Pol 2013;60:183-189.
[11] Delaye L, Becerra A, Lazcano A. The last common ancestor: what's in a name? Orig Life Evol Biosph 2005;35:537–554.
[12] Polanco C, Buhse T, Samaniego JL, Castañón González JA. A toy model of prebiotic peptide evolution: the possible role of relative amino acid abundances. Acta Biochim Pol 2013;60:175-182.
[13] Isaev A. Introduction to Mathematical Methods in Broinformatics Springer-Verlag 2004.
[14] Vidyasagar M. The complete realization problem for Hidden Markov Models: A surveys and some new results. Mathematics of Control, Signals and Systems 2011;23:1-65.
[15] Ching W, Ng M, Fung E. Higher order Hidden Markov Models with applications to DNA secuences. Lectures Notes in Computer Science Springer-Verlag 2003;2690:535-539.
[16] Elliott RV, Aggoun L, Moore JB. Hidden Markov Models. Estimations and Control Springer-Verlag 1995.
[17] Fine S, Singer Y, Tishby N. The Hierarchical Hidden Markov Model: Analysis and Applications. Machine Learning 1998; 32:41-62.
[18] Rabiner LA. Tutorial on Hidden Markov Models and selected applications in speech recognition Proc IEEE 1989;77:257-286.
[19] Nguyen NT, Phung D, Venkatesh S. Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Model Proc. IEEE, 2005;2005:995-960.
[20] Bui H, Phung D. Venkatesh,S. Hierarchical Hidden Markov Models with General State Hierarchy AAAI 2004;324-329.
[21] Polanco C, Buhse T. (2017) Non-racemic mixture model: a computational approach. Acta Biochim Pol 64, 17-19.
[22] Kirk DB, Hwu Wen-Mei W. (2016) Programming Massively Parallel Processors: A Hands-on Approach. Morgan and Kaufmann Publishers. 3rd. Edition.
[23] Patterson D, Hennessy J. (2016) Computer Organization and Design. The Hardware/Software Interface, ARM Edition. Morgan and Kaufmann Publishers.
[24] David B. Kirk , Wen-mei W. Hwu. (2016) Programming Massively Parallel Processors, Third Edition: A Hands-on Approach. 3rd Edition. Morgan Kaufmann.
[25] Duane Storti, Mete Yurtoglu. (2015) CUDA for Engineers: An Introduction to High-Performance Parallel Computing. Addison-Wesley Professional.
[26] Nicholas Wilt. (2018) The CUDA Handbook: A Comprehensive Guide to GPU Programming. 2nd Edition. Addison-Wesley Professional .
[27] Pospichal, Petr & Jaros, Jiri & Schwarz, Josef. (2010). Parallel Genetic Algorithm on the CUDA Architecture. Proceedings of Applications of Evolutionary Computation, EvoApplicatons 2010; 442-451.