OPEN

Three dimensional structures of putative, primitive proteins to investigate the origin of homochirality

Akifumi Oda[1,2,3], Tomoki Nakayoshi[1,3], Koichi Kato[1,4], Shuichi Fukuyoshi[3] & Eiji Kurimoto[1]

Primitive proteins are likely to have been constructed from non-enzymatically generated amino acids, due to the weak enzymatic activities of primitive biomolecules such as ribozymes. On the other hand, almost all present proteins are constructed only from L-amino acids. Therefore, there must have been a mechanism early in the origins of life that selected for one of the optical isomers of amino acids. In this study, we used molecular dynamics simulations to predict the three-dimensional structures of the putative primitive proteins constructed only from glycine, alanine, aspartic acid, and valine ([GADV]-peptides). The [GADV]-peptides were generated computationally at random from L-amino acids (L-[GADV]-peptides) and from both L- and D-amino acids (DL-[GADV]-peptides). The results indicate that the tendency of secondary structure formation for L-[GADV]-peptides was larger than that for DL-[GADV]-peptides, and L-[GADV]-peptides were more rigid than DL-[GADV]-peptides. These results suggest that the proteins with rigid structure motifs were more prone to have been generated in a primordial soup that included only L-amino acids than a the soup including racemic amino acids. The tendency of the rigid structure motif formation may have played a role in selecting for the homochirality that dominates life on Earth today.

In the Hadean eon (about 4,000 million years ago), the first life was generated on Earth. Although the specific details of the origin of life are unclear, several hypotheses about the components of primitive life have been proposed. For example, in the RNA world hypothesis, the first life was presumed to be constructed from RNA[1]. On the other hand, the protein world hypothesis states that life was considered to originate from proteins[2,3]. The protein world hypothesis is one of the "metabolism-first" scenarios in which self-sustaining chemical reactions are considered to be the ancestor of what we call life. In such hypotheses, primitive biomolecules are presumed to act as catalysts for metabolic reactions. The history of metabolism-first hypotheses dates back to Oparin's coacervate theory[4]. Various other metabolism-first scenarios have been proposed such as early chemical evolution theories by Haldane[5], iron-sulfur world[6], garbage-bag world[7], and protein world. In the GADV hypothesis, the proteins constructed from glycine (G), alanine (A), aspartic acid (D), and valine (V) ([GADV]-proteins) were the primary components of primitive life[2,3]. In addition, because G, A, D, and V can be synthesized easily from inorganic compounds under primitive Earth conditions without enzymes, [GADV]-proteins are considered the important biomolecules not only for the protein world hypotheses but also for the RNA world hypothesis[8]. Although proteins today include around 20 types of amino acids, many researchers have considered that primitive proteins were constructed from a limited set of amino acids[2,3,8–11]. The [GADV] amino acid set includes both hydrophobic and hydrophilic amino acids, and all four amino acids are coded for by the universal code GNC. This indicates that these four amino acids can generate proteins with various properties by a simplified genetic code.

Recently, we have predicted the three-dimensional (3D) structures of randomly generated [GADV]-proteins by using computational chemical methods[12]. The results of these studies suggest that even short and randomly generated peptide chains, including only G, A, D and V, can form secondary structures that might act as proteins. On the other hand, peptides used in those studies include only L-amino acids, but D-amino acids were not

[1]Meijo University, Faculty of Pharmacy, Nagoya, 468-8503, Japan. [2]Osaka University, Institute for Protein Research, Suita, 565-0871, Japan. [3]Kanazawa University, Faculty of Pharmacy, Institute of Medical, Pharmaceutical and Health Sciences, Kanazawa, 920-1192, Japan. [4]Kinjo Gakuin University, Faculty of Pharmacy, Nagoya, 463-8521, Japan. Correspondence and requests for materials should be addressed to A.O. (email: oda@meijo-u.ac.jp)

considered. If the amino acids were constructed by non-enzymatic reactions on the primitive Earth in the Hadean eon, the amino acids must have existed in a racemic form. Even researchers who advocate for an RNA world hypothesis agree that non-enzymatically generated amino acids are important components of life because of the weak enzymatic activity of ribozymes. Miller reported that several racemic amino acids were constructed under the putative primitive Earth condition[13]. After the Miller experiment, several researchers have reported that various amino acids can be generated non-enzymatically in a putative primitive Earth environment[14–21]. Although the Miller experiment was performed in a strongly reducing atmosphere, amino acids also were generated under a weakly reducing atmosphere[14]. In addition, amino acids generated in hydrothermal vents[15–18] and in space[19–21] have been found. Thus, many researchers have considered that non-enzymatically constructed amino acids likely existed on the primitive Earth and are the important components of primitive life. Because non-enzymatically generated amino acids reported previously all were racemic[13–21], both L- and D-form of amino acids should be used in the structural predictions of the primitive proteins.

In the proteins in extant life, amino acid residues are almost always in the L-form. Although stereoinversion of the amino acid residues have been found in tissues of aged individuals[22], wild type proteins are almost always constructed from L-amino acids. To date, various hypotheses have been proposed to explain the origin of homochirality of amino acid residues[23–29]. Due to non-enzymatic synthesis, racemic amino acids are generated. Therefore, the inequality in the abundance ratio between L- and D-amino acids was caused at the earliest stages in the origin of life, and the chiral bias was amplified during the course of evolution. Several experimental and computational results have been reported to explain this chiral bias. For example, an enantiometric excess in meteoritic amino acids[23], chiral bias from amino acid degradation of circular polarization[24], from the adsorption on crystal surfaces[25], and from the parity-violating energy shifts of amino acids[26] have been reported. Thus, some hypotheses for the amplification of chiral bias have been proposed, such as enantioselective crystallization[27] and enantioselective peptide formation on water[28]. Furthermore, solubility-based amplification of small enantiometric excess has been reported for amino acids as well as nucleosides[29]. Although these hypotheses explain the physical environments and chemical reactions of amino acids and peptides, biological properties of primitive proteins and peptides were not discussed sufficiently. In particular, the structural biological properties of primitive proteins and peptides constructed from homochiral or racemic amino acids have not been investigated. If the structural properties of proteins constructed from homochiral amino acids were different from those of proteins that include both L- and D-amino acids, the origin of homochirality may be explained by the protein property itself without any special external environment. Similarly, the amplification of chiral bias may be explained by the biological properties of primitive proteins. Therefore, investigations of the structural properties of proteins, including both L- and D-amino acids, are expected to play important roles to clarify only L-amino acids were selected during early life, rather than racemic amino acids. Recently, the structural features of proteins constructed from a mixture of L- and D-amino acid residues have been reported[22,30–37]. Several structural changes have been observed in proteins including D-amino acid residues. For example, the aggregation and precipitation of amyloid β (Aβ), elastin, and α-crystallin have been reported[22]. For Aβ, the formation and collapse of β-sheet structures caused by D-amino acids have been observed using circular dichroic spectra[30,31] and computational chemistry methods[32]. The effects of D-amino acid residues on 3D structures have been investigated not only for natural proteins and peptides but also for artificial ones. The introduction of D-Ala destabilizes a right-handed α-helical coiled coil construct[33]; however, alternating D, L polypeptides were found in which specific types of local structures were stabilized[34]. For artificial peptides, destabilization of the helix structures depended on the types[35] and positions[36] of D-amino acids. Both local changes[37] and whole structural changes[22,30] were observed upon the introduction of D-amino acid residues. This pro- and anti-structural effect of L, D mixtures prompted our study to clarify the effects of D-amino acids in tentative primitive proteins.

In this study, 3D structures of [GADV]-peptides constructed from racemic amino acids were predicted using molecular dynamics (MD) simulations. Although short peptides cannot form rigid 3D structures, their structural tendency can be evaluated. Recently, we tested the MD simulations for structural predictions of small proteins. The results revealed that structure formation tendencies of small proteins, such as the secondary structure formation tendency, can be predicted by MD simulations[38]. Therefore, we conducted structural predictions of putative primitive proteins using MD simulations. Using the predicted structures, we then investigated various structural properties, including the frequencies of secondary structure formation and hydrogen bonding frequencies, to clarify the influence of D-amino acids in the primitive proteins.

Results and Discussion

In this study, the peptides that included only L-amino acids are referred to as "L-[GADV] peptides," and the peptides including both D- and L-amino acids are referred to as the "DL-[GADV] peptides." We conducted replica exchange MD (REMD) simulations[39] of 100 L-[GADV] peptides and 100 DL-[GADV] peptides. The amino acid sequences of these peptides are shown in Table S1. In Fig. 1a, the secondary structure contents calculated from MD trajectories were compared between L-[GADV] and DL-[GADV] peptides. The average secondary structure contents were 0.49 and 0.41 for L-[GADV] peptides and DL-[GADV] peptides, respectively. As shown in the figure, more secondary structures were observed for L-[GADV] peptides than DL-[GADV] peptides even though the amino acid sequences were the same. In addition to the secondary structure contents, when residues formed secondary structures at 50% of the trajectory, the residue was defined as a "structured residue." The number of structured residues indicates the ability to form structures of [GADV]-peptides. The 100 peptides were classified by the number of structured residues (Fig. 1b). Although the maximum number of structured residues was eight for L-[GADV] peptides, at most only six residues can form secondary structure for DL-[GADV] peptides. In addition, 61 DL-[GADV] peptides had no residues that formed secondary structures, whereas 46 L-[GADV] peptides had no structured residues. The average number of structured residues in the 100 peptides were 1.80 and 0.83 for L-[GADV] and DL-[GADV] peptides, respectively. These results indicate that secondary structures are more
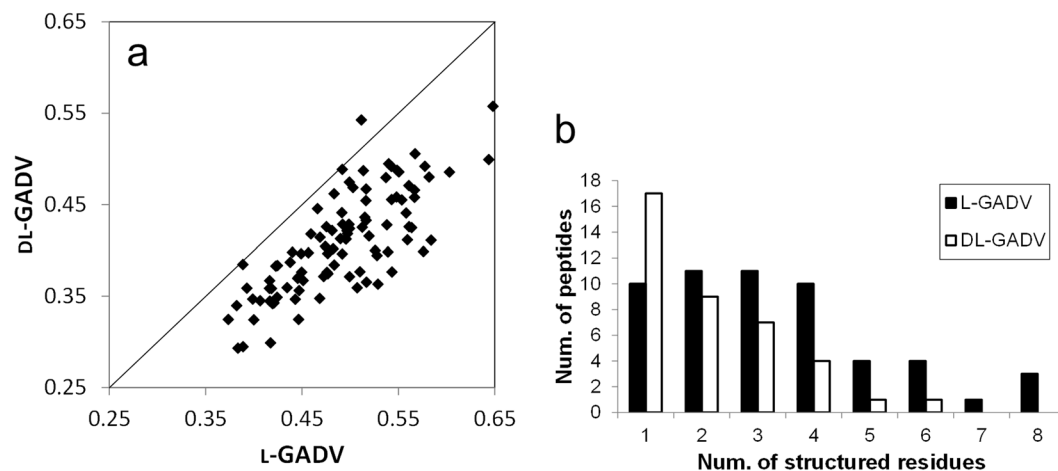
Figure 1. Secondary structures in [GADV] peptides. (**a**) The secondary structure contents. Vertical and horizontal axes indicate the secondary structure contents throughout the last 10 ns trajectories. Each point represents one sequence. The line represents the 1:1 line, where secondary structures are equal for both types of peptides. (**b**) The number of peptides classified by the number of the structured residues. When the residue formed secondary structure at 50% of trajectory, the residue was defined as "structured residue."

prone to form in L-[GADV] peptides than in DL-[GADV] peptides. Because the amino acid sequences were the same for L-[GADV] peptides and DL-[GADV] peptides, these results can be explained by diastereomerism. That is, the inclusion of D-amino acids decreases the ability of secondary structure formation in [GADV]-peptides.

When the helix and β structures were analyzed separately, L-[GADV] peptides were more prone to form a helix rather than a β structure (1.76 vs. 0.04). On the other hand, the occurrence of helical and β residues were approximately the same for DL-[GADV] peptides (0.42 vs. 0.41). As in the previous paper[12], β turns were observed as β structures, and sheet structures were not detected. Of the 200 L-[GADV] and DL-[GADV] peptides, only one peptide included three consecutive β residues (D-Asp12-D-Val13-L-Val14 in peptide 58 of DL-[GADV] peptides), and these three consecutive residues all formed a β turn. In the three residues, alternation between D- and L-amino acids existed (D-Val13-L-Val14). The two consecutive β residues were found in seven DL-[GADV] peptides. In six of the seven DL-[GADV] peptides (including two consecutive β residues), D- and L-forms were bound in consecutive β structures. On the other hand, no L-[GADV] peptides included two consecutive β structures. These results indicate that the bond between D- and L-amino acids is prone to form β structure. However, because only three β residues were found in DL-[GADV] peptides, β structures were not sufficient to explain the rigid protein structure formations. Therefore, although the number of β residues was larger in DL-[GADV] peptides than in L-[GADV] peptides, L-[GADV] peptides seemed more prone to form rigid proteins. As mentioned above, the average number of helix residues for L-[GADV] peptides was more than four times that for DL-[GADV] peptides. By contrast, β-sheet structures were not found in all L-[GADV] and DL-[GADV] peptides. The average β-sheet contents for L-[GADV] and DL-[GADV] peptides were 0.01 and 0.02, respectively. Although occasional exceptions were observed (such as chignolin[40]), helices play an important role in structure formation of the small protein. Because the sequence lengths of the non-enzymatically generated primitive proteins were short, the L-[GADV] peptides that more easily form helix structures appeared to have an advantage in forming proteins under the primitive Earth environment compared to DL-[GADV] peptides.

As shown in Fig. 1, more secondary structures were observed in L-[GADV] peptides than in DL-[GADV] peptides. However, the concept of secondary structure was defined from the analytical results of existing proteins. Thus, proteins that included atypical amino acids, such as D-amino acids, might form atypical secondary structures that cannot be detected by DSSP. Because the rigidity of the amino acid chain plays an important role in protein function, helix, and β structures, as well as atypical secondary structures, may act as functional motifs of proteins. The structure of the catalytic site plays an important role not only in enzymes but also in simpler molecular catalysts[41,42]. When a molecule acts as a catalyst, the catalytic moiety ("active site") must be located near the reaction point of the reactant and an appropriate reactant complex must be formed. The appropriate structure of the reactant complex facilitates a chemical reaction[43,44]. Because the locations of catalytic moieties have to be fixed, a certain level of rigidity plays an important role in primitive proteins that can act as primitive enzymes. Therefore, to investigate the rigidity of [GADV]-peptides, we calculated the RMSF of Cα atoms, which indicates the structural flexibility of Cα atoms. Smaller RMSF values indicate more rigid structures of the atoms. In Fig. 2a, the average RMSFs of Cα atoms were compared between L-[GADV] and DL-[GADV] peptides. The average of the average RMSFs were 4.64 and 4.84 for L-[GADV] peptides and DL-[GADV] peptides, respectively. As shown in the figure, L-[GADV] peptides were more rigid than DL-[GADV] peptides even though the amino acid sequences were the same. In addition to the average RMSFs, the numbers of rigid residues were evaluated. In these calculations, residues with RMSF < 4 Å were defined as "rigid" residues. The numbers of the peptides classified by the number of the rigid residues were illustrated in Fig. 2b. Peptides that included 13 rigid residues were found in L-[GADV] peptides, although at most, nine rigid residues were observed in DL-[GADV] peptides. Only
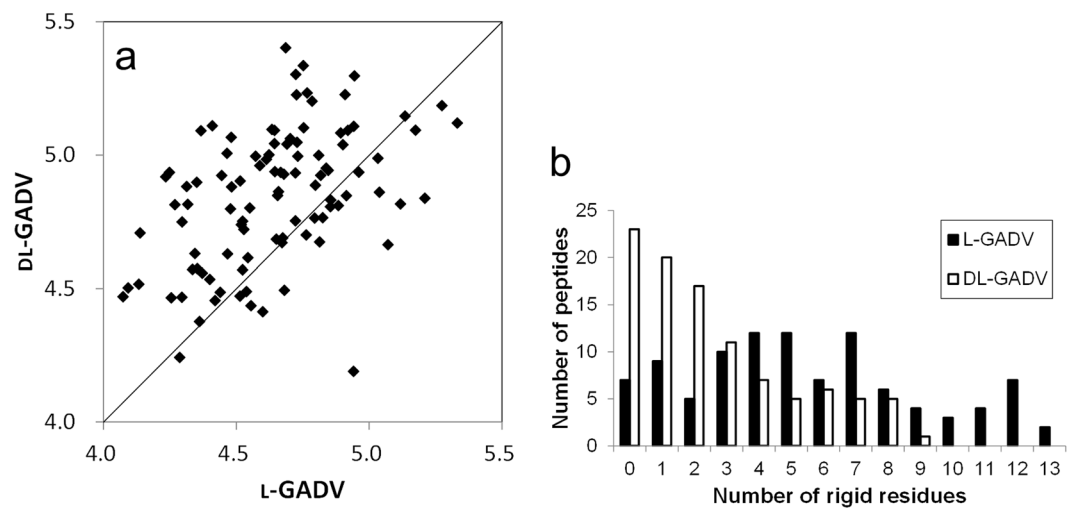
Figure 2. Rigid residues in [GADV] peptides. (**a**) Average RMSF of Cα. Vertical and horizontal axes indicate the average RMSFs of Cα throughout the last 10 ns trajectories. Each point represents one sequence. The line represents the 1:1 line, where RMSF are equal for both types of peptides. (**b**) The number of peptides classified by the number of the rigid residues. The residue in which RMSF of Cα was smaller than 4.0 Å was defined as rigid residue.
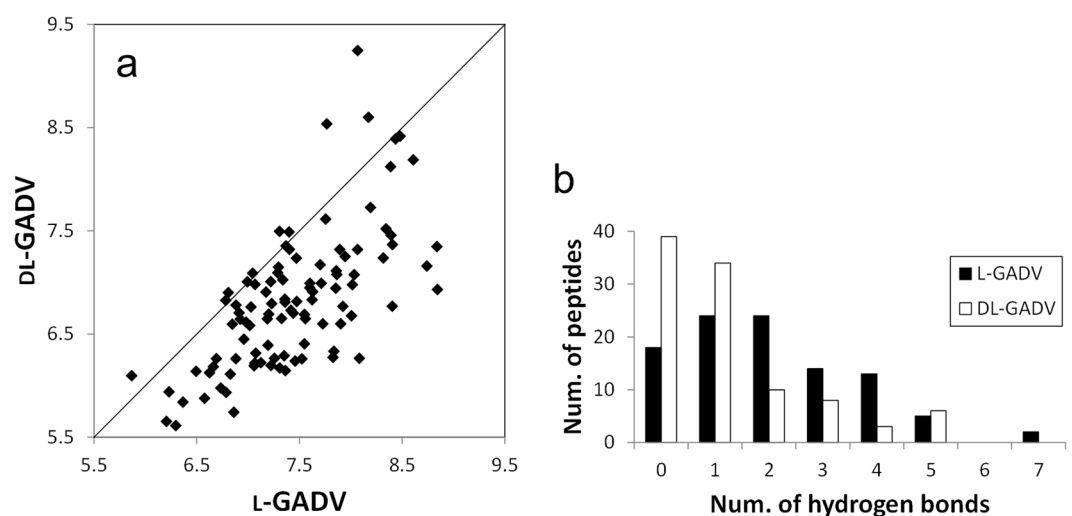

Figure 3. Numbers of the hydrogen bonds. (**a**) Average number of the hydrogen bonds. Vertical and horizontal axes indicate the average numbers of hydrogen bonds throughout the last 10 ns trajectories. Each point represents one sequence. The line represents the 1:1 line, where numbers of bonds are equal for both types of peptides. (**b**) The number of peptides classified by the number of the frequently observed hydrogen bonds. Hydrogen bonds that were observed in 30% of the snapshots in the last 10 ns trajectories were defined as frequently observed hydrogen bonds.

seven L-[GADV] peptides had no rigid residues, but 23 DL-[GADV] peptides had no rigid residues. The average number of rigid residues was 5.51 and 2.60 for L-[GADV] peptides and DL-[GADV] peptides, respectively. In addition, average RMSFs of all residues (20 residues × 100 peptides = 2000 residues) were 4.64 Å and 4.84 Å for L-[GADV] peptides and DL-[GADV] peptides, respectively. These results indicate that the L-[GADV] peptides are more rigid than DL-[GADV] peptides. In other words, the structure of peptides becomes flexible if D-amino acids are included. Therefore, L-[GADV] peptides tend to form normal secondary structures, such as α helix and β structures, as well as other ordered structures, in comparison with DL-[GADV] peptides. These results suggest that L-[GADV] peptides are more "protein-like" than DL-[GADV] peptides.

In Fig. 3a, the trajectory average of the numbers of hydrogen bonds in each peptide is compared between L-[GADV] and DL-[GADV] peptides. Even though the amino acid sequences were the same, more hydrogen bonds were observed for L-[GADV] peptides than DL-[GADV] peptides (Fig. 3a). The average for the 100 trajectory averages of the numbers of hydrogen bonds were 7.44 and 6.82 for L-[GADV] peptides and DL-[GADV] peptides, respectively. These results suggest that hydrogen bonds become difficult to form if D-amino acids are
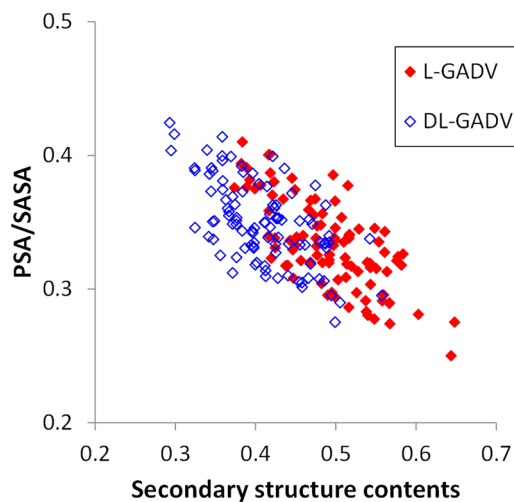
Figure 4. Relationship between PSA/SASA and secondary structure contents. The ratio between the trajectory averages of PSA and SASA was defined as PSA/SASA. The secondary structure content is the ratio of the number of structured residues to the number of all residues, and it is compatible with the secondary structure content obtained by experiments such as circular dichroic spectrum. Each point represents one peptide, and there are 100 red dots and 100 blue dots corresponding to the 100 L-[GADV] and 100 DL-[GADV] peptides, respectively.

included in the tentative primitive proteins. Because the amino acid sequences were the same, the correlation coefficient of the average numbers of hydrogen bonds between L-[GADV] peptides and DL-[GADV] peptides was high (0.670). On the other hand, despite having the same sequences, the number of hydrogen bonds decreased by including D-amino acids. The decreased number of hydrogen bonds may affect the difficulty of rigid structure formation of tentative primitive [GADV]-proteins, including D-amino acids.

In Fig. 3a, all snapshots in the last 10 ns trajectories were used for analyses. Therefore, even rarely observed hydrogen bond that were found in only a few snapshots were counted. However, continuously formed hydrogen bonds play more important roles in protein structure formation than rarely observed hydrogen bonds. Therefore, frequently observed hydrogen bonds were analyzed separately. Hydrogen bonds that were observed in 30% of the snapshots in the last 10 ns trajectories were defined as "frequently observed hydrogen bonds." The number of frequently observed hydrogen bonds is expected to act as an indicator of structure formation tendency. In Fig. 3b, the numbers of the peptides classified by the number of the frequently observed hydrogen bonds were illustrated. For L-[GADV] peptides, not only the average number of hydrogen bonds shown in Fig. 3a but also the numbers of the frequently observed hydrogen bonds shown in Fig. 3b were larger than for DL-[GADV] peptides. There were 18 and 39 peptides in which there were no frequently observed hydrogen bonds for L-[GADV] peptides and DL-[GADV] peptides, respectively (Fig. 3b). In fact, more than 70% of DL-[GADV] peptides included only one or zero frequently observed hydrogen bonds. The average number of the frequently observed hydrogen bonds was 2.05 and 1.20 for L-[GADV] peptides and DL-[GADV] peptides, respectively. These results indicate that the important hydrogen bonds, which are formed continuously to maintain the rigid protein structure, are more difficult to form in DL-[GADV] peptides.

For all snapshots of peptides, PSA and SASA were calculated. PSA/SASA was negatively correlated with secondary structure content (Fig. 4). The correlation coefficients were −0.721 and −0.616 for L-[GADV] peptides and DL-[GADV] peptides, respectively. This result suggests that the more hydrophobic peptides can form more secondary structures, and the hydrophobic effects play important roles in structure formations of proteins. Although PSA/SASA values of DL-[GADV] peptides were slightly higher than for L-[GADV] peptides, the difference was not substantial. The average PSA/SASA for the 100 peptides was 0.333 and 0.349 for L-[GADV] peptides and DL-[GADV] peptides, respectively. On the other hand, the difference in the average value of secondary structure contents was large (0.491 vs. 0.409 for L-[GADV] peptides and DL-[GADV] peptides, respectively). This result indicates that the ability of secondary structure formation for L-[GADV] peptides is superior to that for DL-[GADV] peptides, even if the water solubilities are similar. Because primitive life likely originated in water environments, L-[GADV] peptides appear to be more suitable for primitive proteins, because of a fine balance between water solubility and secondary structure formation ability. Especially for helix contents, the average value of L-[GADV] peptides was larger than that of DL-[GADV] peptides (0.253 for L-[GADV] peptides and 0.163 for DL-[GADV] peptides). Therefore, L-[GADV] peptides seem to be more prone to form helical proteins compared to DL-[GADV] peptides. This study compared L-amino acids and the racemic mixture. The AMBER force field parameters[45,46] for the D-amino acids used in this study are the same as those for the L-amino acids — chirality being enforced by the relative starting positions of atoms[47]. Therefore, by definition, there is no difference between all-L and all-D systems in these MD simulations. The point of this study was not to justify the choice of L- over D-amino acids, but to investigate homochiral vs. racemic amino acids. Therefore, our results do not

explain an all-L world, specifically, but a homochiral world. The results of the present study indicate that homo-chiral [GADV]-peptides are more prone to forming stable structures compared to racemic [GADV]-peptides.

Conclusion

In this study, the 3D structures were investigated with REMD simulations for two types of [GADV]-peptides, one constructed only from L-amino acids and the other that included both L- and D-forms. The results showed that L-[GADV] peptides are more prone to form the rigid and well-ordered structures than DL-[GADV] peptides. Therefore, the "primordial soup" that excluded D-amino acids seems to be more appropriate for an environment favoring the origin of proteins than a soup that included both L- and D-amino acids in the GADV hypothesis. This study compared L-amino acids and the racemic mixture, not L- and D-amino acids. In fact, Towse *et al.* reported that artificial peptides constructed from only D-amino acids can form appropriate conformations as well as peptides consisting of only L-amino acids[48]. Our results suggest that the homochiral primordial soup, which included only L- or D-amino acids, is more preferable as the birthplace of primitive proteins than racemic soup for [GADV]-peptides.

As mentioned above, although the non-enzymatically synthesized amino acids are racemic, many researchers have considered that the chiral bias was caused in the primordial soup by other reasons, such as cosmic rays, parity-violating energy shifts, or mere chance. Because extant proteins are almost all constructed from L-amino acids, the homochirality must have been achieved in the very earliest stages of the origin of life. Our results suggest that the primitive [GADV]-proteins were more prone to originate in a chiral-biased primordial soup than in a racemic soup. In addition, in biological systems, rigid, and structured [GADV]-proteins are more prone to be constructed when only L-amino acids are generated. Because protein production is the key for reproduction, biological systems that can generate proteins more easily seem to reproduce more easily. That is, 3D structure formation of proteins may act as a selection pressure, and the biological system that can generate only L-amino acids may be the one selected. By contrast, systems that generated racemic amino acids were inferior, in terms of [GADV]-protein formation ability, and the inferiority of the protein production may cause delays in the origin of life in the racemic environment. Thus, a racemic system may have been outcompeted, in favor of a chiral-biased system. As future development of this study, structural predictions and hydrogen bonding analyses using the resulting structures of this study in the explicit solvent are needed. To investigate protein quaternary structure, such as amyloid-forming ability, multi-chain simulations should be carried out. The structural formation tendencies of other combinations of amino acids than [GADV]—especially amino acid sets including non-proteogenic amino acids like isovaline and β-alanine—are currently being evaluated. The results of this study indicate that peptides constructed only by L-amino acids are more prone to evolve into complex proteins, which may be one reason for the origin of homochirality of life.

Methods

[GADV]-peptides. Similar to our previous study[12], we used a sequence length of 20 for the [GADV]-peptides. In the peptides, amino acid sequences were determined randomly, and each residue was stereoinverted randomly at a 50% probability. The occurrence rates of G, A, D, and V in the peptides were all 25% (equal rates). A total of 100 sequences were generated for the test. In addition, the peptides including all L-amino acids were also constructed for comparison. To compare between L-[GADV] and DL-[GADV] peptides, we generated the same 100 sequences for both peptide types. For the test peptides, initial 3D structures were generated by the tleap module of AmberTools in the AMBER12 program[49]. To avoid bias from initial structures, linear structures were used for the initial of MD simulations.

Molecular Simulations. First, structural minimizations were performed for the initial linear structures. The maximum number of minimization cycles was set to 2000. After minimization, MD simulations were conducted using the minimized structures. For the structural predictions of [GADV]-peptides, we used REMD[39]. The calculations were conducted using AMBER12[49]. Because we already tested the applicability of the force field parameters for D-amino acids[47], the AMBER ff12SB force field was used for both L- and D-amino acids. The cutoff distance for non-bonding terms was 999 Å, indicating that no cutoff calculations were performed. Water solvent was treated as a generalized Born implicit solvent using a GBneck 2 model (igb = 8)[50] with an mbondi3 atomic radius, and the salt concentration was set at 0.1 M. A total of 50 ns REMD simulations were carried out with a 2 fs time step. Sixteen replicas were used for calculations, and the replica exchange was performed at each 500 steps. The temperatures of replicas were set at 269.7 K, 284.4 K, 300.0 K, 316.4 K, 333.8 K, 352.0 K, 371.3 K, 391.7 K, 413.1 K, 435.7 K, 459.6 K, 484.8 K, 511.3 K, 539.3 K, 568.8 K, and 600.0 K. Before REMD simulations, we performed 200 ps normal MD simulation for each replica to equilibrate the system at each temperature. For all MD simulations, we enforced restrictions to avoid chiral inversions at high temperatures. The interatomic distances of covalent bonds including hydrogen atoms, e.g., hydrocarbon C-H, hydroxyl O-H, and amide N-H, were constrained according to the SHAKE method. Previously, we evaluated the effectiveness of MD simulations for structural predictions of short proteins[38,51]. The results of benchmark studies indicate that the REMD simulation used in this study can be used for secondary structure predictions of small proteins.

Analyses. For the data analyses, the final 10 ns trajectories from a total of 50 ns simulations were used. Because the 3D structures were stored every 1 ps, 10,000 snapshots were included in the 10 ns trajectory. The secondary structures of all snapshots were analyzed, and the secondary structure contents were calculated. The assignments of the secondary structures of proteins were performed with DSSP[52]. For the DSSP results, parallel and anti-parallel sheets and turns were defined as "β structures," and α, 3–10, and π helixes were considered as "helix." The definitions of these secondary structures are identical to those used by refs[12,32,]53. In addition, root mean square fluctuations (RMSF) of Cα atoms were calculated to evaluate the structural flexibilities of

[GADV]-peptides. Because the concept of secondary structure is based on the 3D structures of currently existing proteins, primitive proteins might form unique secondary structures even when they can form well-ordered structures. For example, α strands and β helices were found in the alternating D, L polypeptides[34]. To detect ordered structures other than general secondary structures, RMSF values were calculated. Because the RMSF value indicates the flexibilities of peptides, the rigid portion of the peptides can be detected by forming ordered structures. In addition, the number of hydrogen bonds observed in each snapshot also was counted. The number of intramolecular hydrogen bonds is related to the 3D structure formation. The default settings of the cpptraj module of AmberTools were used to detect hydrogen bonds for all 10,000 snapshots of the last 10 ns REMD trajectory, and the average number of hydrogen bonds across all 10,000 snapshots was defined as the "trajectory average." The trajectory average of the numbers of the hydrogen bonds was calculated for each peptide. Furthermore, the solvent accessible surface area (SASA) and the polar surface area (PSA) of the peptide was calculated using each snapshot. SASA is the molecular surface area that can be accessed by the probe sphere with a 1.4 Å radius. PSA is the surface area of nitrogen and oxygen exposed to solvent water and is used generally as an indicator of hydrophilicity[54]. The RMSF calculations and assignments of hydrogen bonds were conducted with the cpptraj module of AmberTools. SASA and PSA were calculated with the dms program[55,56].

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
2. Ikehara, K. Possible steps to the emergence of life: the [GADV]-protein world hypothesis. *Chem. Rec.* **5**, 107–118 (2005).
3. Ikehara, K. [GADV]-protein world hypothesis on the origin of life. *Orig. Life Evol. Biosph.* **44**, 299–302 (2014).
4. Oparin, A. I. The Origin of Life on the Earth, 3rd ed. (Academic Press, New York, 1957).
5. Haldane, J. B. S. The orgin of life. *The Rationalist Annu.* **3**, 3–10 (1929).
6. Wächtershäuser, G. Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.* **52**, 452–484 (1988).
7. Dyson, F. Origin of Life, 2nd Ed. (Cambridge University Press, Cambrige, 1999).
8. van der Gulik, P., Massar, S., Gilis, D., Buhrman, H. & Rooman, M. The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* **261**, 531–539 (2009).
9. Zuckerkandl, E., Derancourt, J. & Vogel, H. Mutational trends and random processes in the evolution of informational macromolecules. *J. Mol. Biol.* **59**, 473–490 (1971).
10. Jordan, I. K. *et al.* A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**, 633–638 (2005).
11. Tanaka, J., Doi, N., Takashima, H. & Yanagawa, H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* **19**, 786–795 (2010).
12. Oda, A. & Fukuyoshi, S. Predicting three-dimensional conformations of peptides constructed of only glycine, alanine, aspartic acid, and valine. *Orig. Life Evol. Biosph.* **45**, 183–193 (2015).
13. Miller, S. L. A production of amino acids under possible primitive earth. *Science* **117**, 528–529 (1953).
14. Schlesinger, G. & Miller, S. L. Prebiotic synthesis in atmospheres containing CH_4, CO, and CO_2 I. Amino acids. *J Mol Evol* **19**, 376–382 (1983).
15. Hennet, R. J.-C., Holm, N. G. & Engel, M. H. Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: A perpetual phenomenon? *Naturwissenschaften* **79**, 361–365 (1992).
16. Yanagawa, H. & Kobayashi, K. An experimental approach to chemical evolution in submarine hydrothermal systems. *Orig. Life Evol. Biosph.* **22**, 147–159 (1992).
17. Marshall, W. L. Hydrothermal synthesis of amino acids. *Geochim Cosmochim Acta* **58**, 2099–2106 (1994).
18. Huber, C. & Wächtershäuser, G. α-hydroxy and α-amino acids under possible Hadean, volcanic origin-of-life conditions. *Science* **314**, 630–632 (2006).
19. Kobayashi, K., Kaneko, T., Saito, T. & Oshima, T. Amino acid formation in gas mixtures by high energy particle irradiation. *Orig. Life Evol. Biosph.* **28**, 155–165 (1998).
20. Miyakawa, S., Yamanashi, H., Kobayashi, K., Cleaves, H. J. & Miller, S. L. Prebiotic synthesis from CO atmospheres: Implications for the origins of life. *Proc. Natl. Acad. Sci. USA* **99**, 14628–14631 (2002).
21. Weber, A. L. & Miller, S. L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273–284 (1981).
22. Fujii, N. & Saito, T. Homochirality and life. *Chem. Rec.* **4**, 267–278 (2004).
23. Cronin, J. R. & Pizzarello, S. Enantiometric excesses in meteoritic amino acids. *Science* **275**, 951–955 (1997).
24. Bailey, J. *et al.* Circular polarization in star-formation regions: Implications for biomolecular homochirality. *Science* **281**, 672–674 (1998).
25. Soai, K. *et al.* D- and L-quartz-promoted highly enantioselective synthesis of a chiral organic compound. *J. Am. Chem. Soc.* **121**, 11235–11236 (1999).
26. MacDermott, A. J., Fu, T., Hyde, G. O., Nakatsuka, R. & Coleman, A. P. Electroweak parity-violating energy shifts of amino acids: The "conformation problem. *Orig. Life Evol. Biosph.* **39**, 407–437 (2009).
27. Kojo, S. & Tanaka, K. Enantioselective crystallization of D,L-amino acids induced by spontaneous asymmetric resolution of D,L-asparagine. *Chem. Comm.* 1980–1981 (2001).
28. Zepik, H. *et al.* Chiral amplification of oligopeptides in two-dimensional crystalline self-assemblies on water. *Science* **295**, 1266–1269 (2002).
29. Breslow, R. & Cheng, Z. On the origin of terrestrial homochirality for nucleosides and amino acids. *Proc. Natl. Acad. Sci. USA* **106**, 9144–9146 (2009).
30. Sakai-Kato, K., Naito, M. & Utsunomiya-Tate, N. Racemization of the amyloidal β Asp1 residue blocks the acceleration of fibril formation caused by racemization of the Asp23 residue. *Biochem. Biophys. Res. Commun.* **364**, 464–469 (2007).
31. Janek, K. *et al.* Study of the conformational transition of Aβ(1–42) using D-amino acid replacement analogues. *Biochemistry* **40**, 5457–5463 (2001).
32. Oda, A., Kobayashi, K. & Takahashi, O. Comparison of molecular dynamics simulation methods for amyloid β_{1-42} monomers containing D-aspartic acid residues for predicting retention times in chromatography. *J. Chromatogr. B* **873**, 3337–3343 (2011).
33. Fairman, R., Anthony-Cahill, S. J. & DeGrado, W. F. The helix-forming propensity of D-alanine in a right-handed α-helix. *J. Am. Chem. Soc.* **114**(13), 5458–5459 (1992).
34. Durani, S. Protein design with L- and D-α-amino acid structures as the alphabet. *Acc. Chem. Res.* **41**, 1301–1308 (2008).
35. Krause, E., Bienert, M., Schmieder, P. & Wenschuh, H. The helix-destabilizing propensity scale of D-amino acids: The influence of side chain steric effects. *J. Am. Chem. Soc.* **122**, 4865–4870 (2000).

36. Hong, S. Y., Oh, J. E. & Lee, K. Effect of D-amino acid substitution on the stability, the secondary structure, and the activity of membrane-active peptide. *Biochem. Pharmacol.* **58**, 1775–1780 (1999).
37. Rothemund, S. *et al.* Structure effects of double D-amino acid replacements: A nuclear magnetic resonance and circular dichroism study using amphipathic model helices. *Biochemistry* **34**, 12954–12962 (1995).
38. Kato, K., Nakayoshi, T., Fukuyoshi, S., Kurimoto, E. & Oda, A. Validation of molecular dynamics simulations for prediction of three-dimensional structures of small proteins. *Molecules* **22**, 1716, https://doi.org/10.3390/molecules22101716 (2017).
39. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
40. Honda, S., Yamasaki, K., Sawada, Y. & Morii, H. 10-residue folded peptide designed by segment statistics. *Structure* **12**, 1507–1518 (2004).
41. Greeley, J. P. Active site of an industrial catalyst. *Science* **336**, 810–811 (2012).
42. Norskov, J. K. *et al.* The nature of the active site in heterogeneous metal catalysis. *Chem. Soc. Rev.* **37**, 2163–2171 (2008).
43. Foresman, J. & Frisch, A. Exploring Chemistry with Electronic Structure Methods, 3rd ed. (Gaussian, Inc., Wallingford, 2015).
44. Szabo, A. & Ostlund, N. S. Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory (Dover Publications, Inc. Mineola, 1996).
45. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
46. Maier, J. A. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
47. Oda, A. *et al.* Validation of molecular force field parameters for peptides including isomerized amino acids. *Chirality* **30**, 332–341 (2018).
48. Towse, C., Hopping, G., Vulovic, I. & Daggett, V. Nature versus design: the conformational propensies of D-amino acids and the importance of side chain chirality. *Protein Eng. Des. Sel.* **27**, 447–455 (2014).
49. Case, D. A. *et al.* AMBER 12 (University of California, San Francisco, 2012).
50. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized Born solvent model parameters for protein simulations. *J. Chem. Theory. Comput.* **9**, 2020–2034 (2013).
51. Kato, K., Nakayoshi, T., Fukuyoshi, S., Kurimoto, E. & Oda, A. Evaluation of molecular dynamics simulation for small protein structural prediction in different simulation time. *Journal of Research Institute of Meijo University* **17**, 41–50 (2018) (Japanese).
52. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
53. Ikebe, J., Kamiya, N., Ito, J., Shindo, H. & Higo, J. Simulation study on the disordered state of an Alzheimer's β amyloid peptide Aβ(12–36) in water consisting of random-structural, β-structural, and helical clusters. *Protein Sci.* **16**, 1596–1608 (2007).
54. Palm, K., Stenberg, P., Luthman, K. & Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **14**, 568–571 (1997).
55. Richards, F. M. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151–176 (1977).
56. Huang, C. DMS (University of California, San Francisco, 2002).

Acknowledgements

Author Contributions

A.O. and E.K. conceived and designed the research, T.N., K.K. and S.F. carried out the simulations, A.O., T.N. and K.K. analyzed the data, A.O. and T.N. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-48134-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.