On the possible origin of protein homochirality, structure, and biochemical function

Jeffrey Skolnick[a,1], Hongyi Zhou[a], and Mu Gao[a]

[a]Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332

Edited by Eugene V. Koonin, National Institutes of Health, Bethesda, MD, and approved November 13, 2019 (received for review May 13, 2019)

Living systems have chiral molecules, e.g., native proteins that almost entirely contain L-amino acids. How protein homochirality emerged from a background of equal numbers of L and D amino acids is among many questions about life's origin. The origin of homochirality and its implications are explored in computer simulations examining the stability and structural and functional properties of an artificial library of compact proteins containing 1:1 (termed demi-chiral), 3:1, and 1:3 ratios of D:L and purely L or D amino acids generated without functional selection. Demi-chiral proteins have shorter secondary structures and fewer internal hydrogen bonds and are less stable than homochiral proteins. Selection for hydrogen bonding yields a preponderance of L or D amino acids. Demi-chiral proteins have native global folds, including similarity to early ribosomal proteins, similar small molecule ligand binding pocket geometries, and many constellations of L-chiral amino acids with a 1.0-Å RMSD to native enzyme active sites. For a representative subset containing 550 active site geometries matching 457 (2) 4-digit (3-digit) enzyme classification (E.C.) numbers, native active site amino acids were generated at random for 472 of 550 cases. This increases to 548 of 550 cases when similar residues are allowed. The most frequently generated sequences correspond to ancient enzymatic functions, e.g., glycolysis, replication, and nucleotide biosynthesis. Surprisingly, even without selection, demi-chiral proteins possess the requisite marginal biochemical function and structure of modern proteins, but were thermodynamically less stable. If demi-chiral proteins were present, they could engage in early metabolism, which created the feedback loop for transcription and cell formation.

origin of protein chirality | origin of life | early metabolism | metabolism first world | emergence of chiral proteins

One striking feature of biological macromolecules is that they are chiral; for example, proteins mainly contain L-amino acids (1). One of the mysteries of the origin of life is how chiral systems emerged from a background of equal amounts of D and L amino acids (2–5). The RNA world hypothesis conjectures that RNA came first. These chiral molecules stored genetic information and catalyzed chemical reactions (6–10). Alternatively, in a minority view, Dyson conjectured that early, probably protein, molecules evolved at least part of the necessary chemistry of life, viz. metabolism, before transcription emerged (11). However, in both views—replication first or metabolism first—the question remains: how was symmetry broken to yield chiral systems? Turning to proteins, there is evidence that carbonaceous meteorites contain an excess of L over D amino acids, with the relative preference depending on meteorite origin (12–15). This could partly explain proteins' L-chirality, assuming that proteins could be made from short polypeptides. To address this issue, Dill and coworkers recently proposed the foldamer hypothesis whereby short hydrophobic protein chains collapse to compact structures, which then catalyze the formation of longer proteins from shorter ones. This view differs from Lupas (16, 17), who proposed that ancient proteins folded by fusion and recombination from ancestral peptides resulting from RNA-dependent translation and catalysis (18). Alternatively, using the "molecules in mutualism hypothesis," nucleotides and amino acids might have catalyzed the synthesis of both (19). More recent studies suggest that aminonitriles, amino acid precursors, readily form peptides in water, providing another source of non–RNA-based proteins (20). Whatever their origin, the minimal conjecture that protein sequences containing equal amounts of D and L amino acids, termed demi-chiral proteins, were present in the prebiotic "soup" is the starting point of the present analysis, which explores the stability and structural and functional properties of demi-chiral model proteins, 3:1 mixtures of D:L and L:D amino acids and homochiral D and L proteins.

Key Questions About Demi-Chiral Proteins

At first glance, one might imagine that the stability and structural and functional properties of demi-chiral proteins are strikingly different from contemporary homochiral, L-amino acid proteins. Since regular secondary structures form from homochiral sequences of amino acids, in demi-chiral proteins, one might expect that the average length of regular secondary structure elements formed by helical stretches of the same chirality might be shorter. What effect does this have on the ability of demi-chiral proteins to form internal hydrogen bonds? Are demi-chiral proteins inherently less stable than chiral ones because they contain fewer internal hydrogen bonds? Are the folds of demi-chiral proteins different from present ones (21)? For native proteins, the library of solved compact single domain native proteins has been shown to be essentially complete, viz. every native protein structure has statistically significant structural similarity (22–24) to members of a library of randomly generated, artificial compact protein structures (25, 26). But what happens when the lengths of regular secondary structural elements are shorter? Are their global folds different? This would imply a discontinuous structural transition from demi-chiral to

Significance

Living systems contain mainly chiral macromolecules, including proteins. How L-chiral proteins emerged from demi-chiral mixtures is unknown. Our simulations show that, compared to contemporary proteins, demi-chiral proteins have shorter regular secondary structures due to fewer internal hydrogen bonds, but similar global folds and small molecule binding sites. Demi-chiral proteins contain L-chiral substructures matching native active site geometries. Among the most frequently generated enzymes with native active site residues are ancient functions associated with metabolism and replication. This suggests that demi-chiral proteins could engage in early metabolism, creating the feedback loop for transcription and cell formation partly responsible for life's emergence.

Author contributions: J.S. designed research; J.S., H.Z., and M.G. performed research; J.S., H.Z., and M.G. analyzed data; and J.S., H.Z., and M.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: skolnick@gatech.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1908241116/-/DCSupplemental.

First published December 10, 2019.

chiral proteins. How different are the geometries and shapes of their small molecule ligand binding pockets from those in homochiral proteins (27, 28)? If they were very dissimilar, then the fundamental chemistry of such putative early, prebiotic proteins could have been very different from now. At the least, this would have profound implications for the validity of the Dyson model that metabolism came first: metabolism and small molecule-based intermolecular signaling would have to be dramatically modified as the transition to chiral systems occurred. Conversely, if they are similar, then their chemistry could be related, providing circumstantial support for the metabolism-first hypothesis.

We next turn to the key question of whether demi-chiral proteins could catalyze chiral reactions. At first glance, one might say no; after all, the system is globally demi-chiral with, on average, equal numbers of D and L amino acids, so how can it do chiral chemistry? On deeper analysis, in a 1:1 mixture of L and D amino acids, one could, by chance, have a constellation of L amino acids at precisely the correct spots in the protein sequence to recapitulate both native active site geometry and chirality, thereby resulting in low level enzymatic activity (29, 30). By symmetry, another protein with opposite amino acid chirality would catalyze the mirror-image reaction. This would then preserve global demi-chirality. Does this happen in randomly generated demi-chiral proteins? If so, what is the relationship between the most frequently found sequences that adopt active site geometries in the demi-chiral system and the minimal gene sets putatively present in the last universal common ancestor (31, 32). If a positive correlation were found, this suggests that these artificial systems might have captured aspects of early proteins.

Results

In what follows, we examine the stability and structural and functional properties of demi-chiral proteins in detail and then compare them to their more chiral counterparts. As previously, we initially consider a library of artificial proteins composed of leucine side chains. Polyleucine is chosen because, in L-chiral proteins, leucine generates compact protein structures whose global volumes and small molecule ligand binding sites match native proteins (25, 33–36). In the following, we examine the secondary structure properties, hydrogen bond energy, and the relative contributions of the secondary structure propensities, burial and pair energies, of compact demi-chiral D:L, 3D:L, D:3L, and pure D and L artificial proteins. Next, we explore the global structural space of demi-chiral proteins. Are their global folds different from contemporary native proteins? In particular, are the structures of early ribosomal proteins in the D:L protein structural library (37)? We then compare the structural similarity of the 3 largest ligand binding pockets in D:L proteins to native ones. Subsequently, we search the library of compact folds for active site geometries containing L amino acids whose root-mean-square deviation, RMSD, is <1.0 Å to native active sites. We show that, in a representative library of 4,516 D:L structures, the native geometries of 550 distinct active sites corresponding to 457 4-digit enzyme classification (E.C.) numbers and 2 3-digit E.C. numbers are found in 413 distinct, compact D:L structures. Then, for each of the 413 distinct D:L structures, following the previously used procedure (34), we randomly generate a given amino acid composition from a shifted version of native composition frequencies to minimize possible bias. We included all 20 contemporary amino acids rather than the likely most ancient ones (38) to enable comparison to contemporary active sites, but the results are similar if a more restricted set of amino acid types is used. Then, the sequence is randomly permuted using a genetic algorithm and selected for predicted stability based on a statistical potential in the given compact polyleucine structure (additional details in *Materials and Methods* and *SI Appendix*), with the lowest-energy sequence selected for subsequent analysis. We then examine whether the most frequently found active sites from independently generated sequences with the correct active site L-residues correlate with ancient enzymatic functions. Finally, we discuss the possible ramifications of our results for the origin of the biochemistry of life.

Hydrogen Bonding and Secondary Structure in Demi-Chiral Proteins

As indicated in Fig. 1, hydrogen bonding in regular secondary structural elements cannot occur between L and D amino acids. Fig. 2*A* clearly shows that the average hydrogen bond energy in the artificial polyleucine demi-chiral proteins is dramatically less than in the more chiral ones. Fig. 2*B* plots the fraction of proteins with greater than the fraction of hydrogen-bonded residues on the abscissa. For demi-chiral proteins, only half have >30% of their residues hydrogen-bonded. In contrast, in pure D or L proteins (with the same results as expected by symmetry), 60% of their residues have internal hydrogen bonds, while 3D:L or D:3L proteins have ∼45% of their residues with intraprotein hydrogen bonds. Thus, independent of the particular force field used, these artificial D:L proteins should be less stable than homochiral ones.

These results suggest that, due to fewer internal hydrogen bonds, the native conformation of demi-chiral proteins is predicted to be thermodynamically less stable than their more globally chiral counterparts. Greater stability could have been one main driving force toward more chiral systems. Due to random fluctuations in composition, even if, on average, proteins contain 50% L and 50% D amino acids, some proteins would have an excess of either D or L amino acids. Those with an unequal number of D and L residues would have more stable compact conformations, all else being equal. Functional selection could also have been operative, but it is ignored here to explore the ramifications of stability selection alone. Consistent with the observation that more ancient superfamilies contain more hydrophobic residues (39), to improve the stability of the folded demi-chiral proteins(s), they might have had a more hydrophobic interior consistent with solubility requirements.

SI Appendix, Fig. S3*A* plots the fraction of secondary structure element lengths defined as helix or extended greater than the value on the abscissa. As expected, pure D and L proteins are indistinguishable, with an average secondary structure length of 6.8 residues. Similarly, 3D:L and D:3L proteins have indistinguishable curves with average secondary lengths of 4.63 and 4.65 residues. Finally, for D:L proteins, the average secondary structural length is 2.95 residues. Even these artificial demi-chiral proteins are locally stiff, although they lack long stretches of helical or beta states. *SI Appendix*, Fig. S3 *B* and *C* plots the number of regular secondary structural elements per protein versus the distribution of lengths greater than the abscissa for α-helices and β-strands, respectively. Demi-chiral proteins can have short α-helices. Interestingly, there are very few β-strands in demi-chiral proteins, but, consistent with *SI Appendix*, Fig. S3*A*, they have extended regions without the characteristic β-strand hydrogen bonding pattern. For the 3D:L or D:3L cases, these proteins have significant amounts of regular helical and beta secondary structures.

Assessing the Relative Contribution of Non–H-bond and H-Bond Pair Potentials for Different Mixed D and L Ratios in Demi-Chiral Structures with Recovered Sequences

The energy distributions of the randomly generated, "recovered" low-energy sequences that adopt the given folds are examined next. For D:L proteins, the average energy per residue, the predicted <E>, is −1.93 in KT units, with the ratio of secondary structure:burial:pair energies of 0.35:0.13:0.52. If we consider the most stable D:L proteins whose predicted energy E is <−3.0, their average energy distribution shifts to 0.23:0.10:0.67. Thus, they have more stabilizing tertiary interactions. For 3D:L proteins, we find that the predicted <E> = −3.42 KT, with relative energy ratios of 0.54:0.07:0.39, while D:3L <E> = −3.45 KT, with the same relative energy ratios of 0.54:0.07:0.39. For pure D

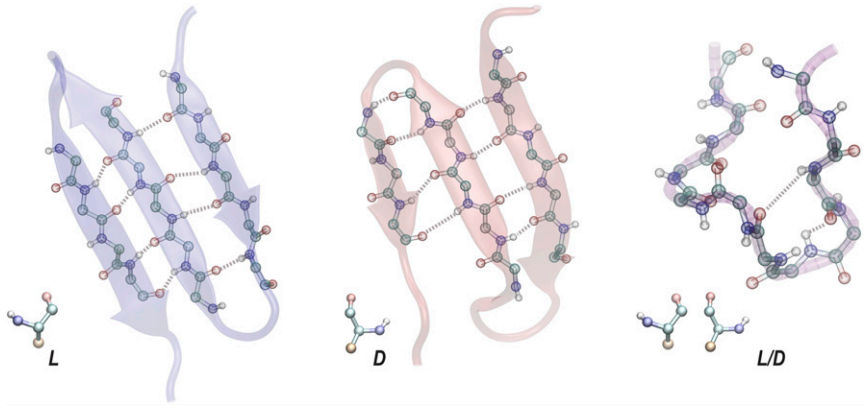

Fig. 1. Schematic representation of (*A*) left-handed, (*B*) right-handed, and (*C*) L-D mixtures of β-strand or backbone extend-state hydrogen bonds. Red indicates the carbonyl oxygen and blue the amide nitrogen atoms associated with backbone hydrogen bonding shown by dashed lines.

proteins, the predicted <E> = −3.70 KT, with the corresponding ratios of energies of 0.50:0.08:0.42. For pure L proteins, the predicted <E> = −3.63 KT, whose energetic ratios are 0.49:0.08:0.43. Thus, these artificial demi-chiral systems are much less energetically stable, and, because they have shorter secondary structure elements, their burial and pair energies are proportionately more important. However, even when 25% of a protein's residues are of opposite chirality, the energy distribution is close to that in homochiral systems, and, importantly, their average energy per residue is ∼94% of homochiral proteins. As the asymmetry in chiral composition increases, proteins rapidly become far more stable. This could act as a thermodynamic driving force toward homochiral systems.

To further confirm our conclusion that folded demi-chiral proteins are predicted to be less stable due to fewer internal hydrogen bonds, we also used 2 other different potentials, DFIRE (used in selecting compact structures) and the Rosetta ab initio force field (40). Both are widely used in protein structure prediction and refinement and have excellent performance in the CASP experiments (40–42). For pure L, 3L:1D, L:D, 1L:3D, and pure D structures, the DFIRE potential gives the ratios of non–H-bond:H-bond pair potentials of 0.40:0.60, 0.49:0.51, 0.60:0.40, 0.49:0.51, and 0.40:0.60. The Rosetta ab initio pair potential has almost identical ratios of 0.40:0.60, 0.49:0.51, 0.60:0.40, 0.49:0.51, and 0.39:0.61 for pure L, 3L:1D, L:D, 1L:3D, and pure D structures, respectively.

For pure L proteins, the average total energy (pair plus H-bond energy) per residue is −1.71 KT using DFIRE and −1.68 KT using the Rosetta ab initio potential. For pure D proteins, the average total energy (pair plus H-bond energy) per residue is −1.69 KT using DFIRE and −1.65 KT by Rosetta ab initio. For D:3L proteins, the average total energy (pair plus H-bond energy) per residue is −1.25 KT using DFIRE and −1.22 KT by Rosetta ab initio. For 3D:L proteins, the average total energy (pair plus H-bond energy) per residue is −1.25 KT using DFIRE and −1.22 KT by Rosetta ab initio. For demi-chiral proteins, with D:L, the average total energy (pair plus H-bond energy) per residue is −1.00 KT using DFIRE and −0.98 KT by Rosetta ab initio. Again, the results are virtually the same whether DFIRE or Rosetta ab initio potentials are used.

Just as was the case earlier where an independent statistical potential was used to generate the putatively fit sequences in the demi-chiral structures, demi-chiral proteins are predicted to be less stable than their more chiral counterparts, with a smaller relative H-bond contribution. The ratio of the average predicted protein stability for D:L to pure D for the statistical potential used in sequence selection is 0.53, whereas DFIRE and Rosetta ab initio suggest that this ratio is 0.59; i.e., it is qualitatively the same. Thus, we recover the same qualitative trends using 3 independent force fields. This strongly suggests that the conclusions are quite general.

Demi-Chiral Proteins Have Native Folds. For the artificial demi-chiral polyleucine structural library, Fig. 3 shows the cumulative fraction of the best Protein Data Bank (PDB) protein structural alignment generated by the structural alignment algorithm

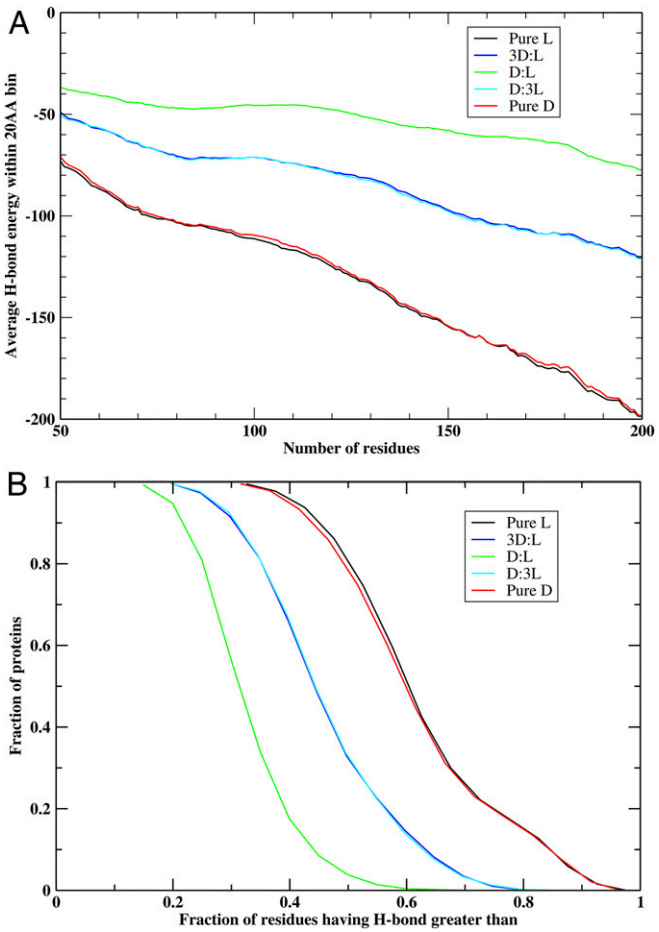

Fig. 2. For pure L, pure D, D:3L, 3D:L, and D:L proteins, (*A*) average hydrogen bond energy within a 20-residue sliding window per protein calculated over proteins vs. the number of protein residues and (*B*) fraction of proteins with at least the fraction of hydrogen-bonded residues on the abscissa.

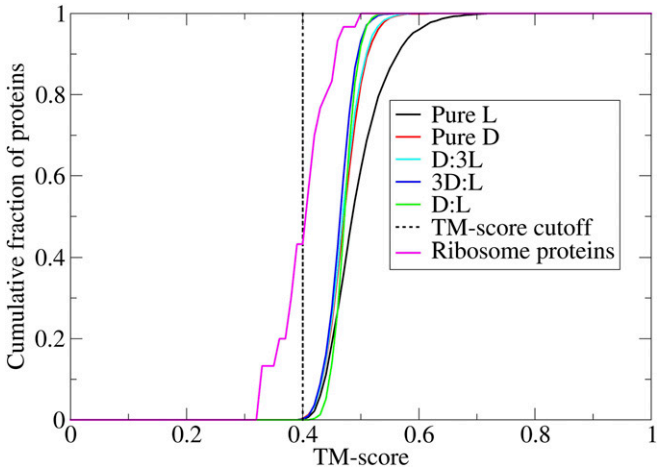

Fig. 3. Cumulative fraction of proteins whose best TM-score is less than or equal to the value on the abscissa obtained from aligning the representative PDB library to the 4,516 protein structures in the pure L, pure D, D:3L, 3D:L, and D:L structural libraries. The purple curve is the best TM-score distribution of the lowest- and next lowest-energy demi-chiral structures to 33 universal ribosomal proteins. The TM-score cutoff for significant fold similarity of 0.4 is shown as the vertical dashed line.

TM-align to the model demi-chiral proteins whose TM-score (22–24) is no more than the given value for the pure L, pure D, D:3L, 3D:L, and D:L (demi-chiral) proteins. The representative native protein library contains 36,799 proteins clustered at 90% pairwise sequence identity (*SI Appendix*, LIST.pdb90). The TM-score range is [0,1]. A value of 1.0 indicates identical folds, with a score of about 0.3 for 2 randomly related structures. A TM-score >0.4 (indicated as the dotted line) indicates that the 2 folds are structurally very similar if not identical. Structures whose TM-score = 0.4 have a P value of 3.4×10^{-5} (24). At first glance, given that artificial demi-chiral proteins lack beta strands (yet have extended states) and much shorter helices, one might conjecture that their global folds could be different from native proteins. However, it was previously shown that protein chains devoid of secondary structure and hydrogen bonding when randomly packed into a sphere whose radius of gyration is that of native proteins have very similar global folds as native ones (43). Basically, they have a similar geometric arrangement of their atoms with close global chain contours, but differ in the absence of longer, regular hydrogen-bonded secondary structures. Thus, Fig. 3 confirms the expected result that the library of demi-chiral protein structures along with the other sets of varying global chirality matches structures in the contemporary PDB library (44). This is a nontrivial conclusion given the fact that the structures are folded from random conformations using the chunk-TASSER ab initio folding algorithm (45).

Ancient Ribosomal Protein Folds. We next examine how many ancient ribosomal proteins have structurally similar folds in the demi-chiral protein structure library. From ref. 37, the ancient ribosomal proteins are L1-6, L10-16, L18, L22-24,L29, L30, S2-5, S7-15, and S19. Fig. 3 shows the cumulative fraction of proteins (purple curve) whose TM-score is no greater than the abscissa. A total of 60% of the universal ribosomal proteins have a TM-score ≥0.4 to the best structural alignment of a demi-chiral protein. The remainder, L2-6, L13, L15, L16, S2, S3, S5, S10, and S12, either have an unstructured long tail that interacts with RNA in the ribosome or 2 spatially disjointed protein domains much like the open jaws of a Pac-Man (each compact domain has significant structural matches in the demi-chiral structural library), whose conformation is stabilized by ribosomal RNA interactions. Again, the space of proteins is only complete for compact, single-domain protein structures (46, 47). Thus, demi-chiral proteins would have the requisite fold geometry to interact with ribosomal RNA provided that an appropriate protein sequence were present.

Demi-Chiral Proteins Have Native-Like Small Molecule Ligand Binding Pockets. We next compared the structural similarity of the small molecule ligand binding pockets in these artificial demi-chiral polyleucine proteins to native proteins. Pocket comparison was performed using APoc and assessed by its PS-score, the range of which is [0,1] (48). Cavities with a PS-score >0.35 have a P value <0.05, with a score of 1 indicating identical pockets (48). Pockets were detected by Cavitator (48), which is good at identifying similar pockets in low-resolution protein models. As shown in Fig. 4, in every case, for pockets containing ≥10 residues and a volume >100 Å^3, native protein pockets have a PS-score >0.35 to the largest, second-largest, and third-largest pockets in demi-chiral proteins. Note that this is for geometric pockets, which are often larger than the pocket bottom most often involved in ligand binding (49). As the geometric pockets get smaller, the structure similarity of native to DL protein pockets increases.

A total of 85.0% of the 213,100 pockets in the native pocket library have a statistically significant match to the largest pockets in the demi-chiral protein library, while 97.1% and 98.7% of native pockets match the second- and third-largest DL pockets. A total of 99.1% of all native pockets have a statistically significant match to at least one of the top 3 DL pockets. The reason for this high coverage is that the library of native pockets is complete and comprised of roughly 500 distinct ligand binding pockets (34); all result from packing defects between secondary structural elements (43). These observations are significant in that the shapes of the pockets in demi-chiral proteins are essentially the same as in modern proteins. Given appropriate sequences, the ability to bind and possibly do catalysis is present in the demi-chiral protein library. If such proteins were in the prebiotic soup, they could potentially bind the appropriate complement of contemporary ligands, albeit probably weakly. Such ancient proteins could engage in very similar chemistry as contemporary ones provided that the relevant proteins, local chirality, and ligands were present.

Turning to pure D and pure L proteins, native protein pockets match 89.3% and 91.3% of their largest pockets, respectively. For the second- and third-largest pockets, pure L (D) proteins

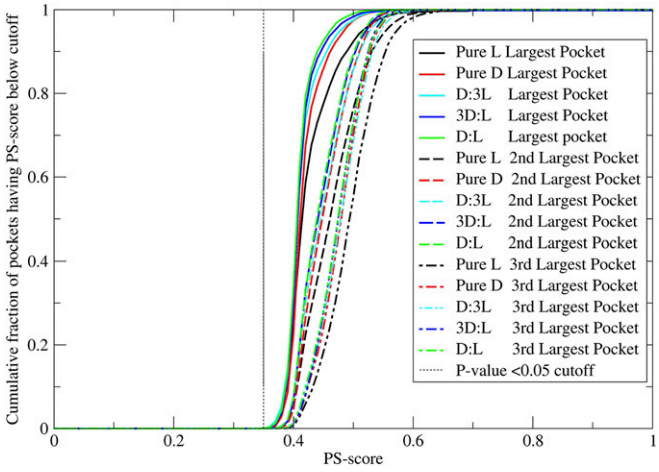

Fig. 4. Plot of the cumulative best PS-score of native pockets to the largest, second-largest, and third-largest pockets in pure L, pure D, D:3L, 3D:L, and D:L protein libraries versus PS-score. The PS-score cutoff with a P value <0.05 is shown as the vertical gray dashed line.

match 94.5% (94.0%) and 99.0% (98.5%) of all native pockets. When all 3 pockets are included, pure L (D) and pure D proteins match 99.4% (98.9%) of all native packets.

The largest pocket in 3D:L (D:3L) proteins match 86.0% (87.8%) of native pockets. For the second- and third-largest pockets, 3D:L (D:3L) proteins match 93.5% (93.8%) and 98.6% (98.8%) of native pockets. Considering all 3 pockets, 99.0% (99.2%) of all native pockets have matches.

These results again point out that protein ligand binding pocket geometry is only weakly dictated by the chirality of the protein backbone and results mainly from defects in secondary structure packing. This is also reflected in the PS-score distribution shown in Fig. 4 as one moves from demi-chiral to pure L proteins. There is a minor effect due to L-chirality that shifts the plateau region from a PS-score of 0.5 to 0.6, as the geometrically finer details and longer secondary structural elements are recovered as the L-amino acid content is increased to the pure L protein case.

Since DL to pure D or pure L protein pockets match essentially all native protein pockets, given the appropriate constellation of amino acids, this implies that proteins of varying global chirality could perform the same chemistry as contemporary native proteins, and there would be no abrupt transition in chemistry from demi-chiral to pure L proteins. In other words, the ability to do chemistry as dictated by protein pocket shape is an inherent protein feature.

Demi-Chiral Proteins Have Native Active Sites Whose Discovery Frequency Correlates with Essentiality. As demonstrated here earlier, demi-chiral proteins possess native-like ligand binding pockets that cover the space of all small molecule ligand binding pockets. As such, we could expect that active site geometries should be found in demi-chiral proteins. In practice, 413 of the polyleucine demi-chiral structures contain active site geometries within a 1.0-Å RMSD from 550 distinct, native active sites. As indicated in *SI Appendix*, Table S1, there are multiple sites hit with the same 4 E.C. numbers. Of the 593 distinct E.C. numbers in the CSA library (50), with the requirement that all L-amino acids be located in the active site geometry within a 1.0-Å RMSD cutoff, 457 of 593 active sites (76%) are hit. If a 3-Å RMSD cutoff is allowed, then 554 of 593 (92%) of all 4 E.C. numbers match. Roughly half of the 49 missing E.C. numbers have NCAT > 5, and 51% are in all 3 domains of life. Removing the L-amino acid requirement slightly increases this ratio to 563 of 593 active sites (95%). Since these geometric requirements are quite stringent and depend on quite fine local structural details, we expect that essentially all E.C. numbers would be hit when additional demi-chiral structures are generated, but this must be established.

As done previously for L proteins, for each polyleucine structure containing a native active site geometry, randomly generated sequences of fixed amino acid composition are then shuffled at random to generate low-energy sequences that putatively adopt that structure (34, 36, 51) (*SI Appendix*, *Materials and Methods*, Stage II). We next select a set of 50 low-energy sequences that fit the given D:L structure and have the appropriate residue types in the active site and assess stability of each sequence in the selected demi-chiral structure. Starting from the initial polyleucine structure on which each sequence is mounted, the structures are relaxed using the chunk-TASSER ab initio folding algorithm. In 75% of cases, the best of the top 5 predicted structures have a TM-score >0.4 to the initial DL structure. Among them, 95 distinct protein D:L fold/sequence pairs with a TM-score >0.8 to the initial DL structure are found. These correspond to close homology models (*Materials and Methods* and *SI Appendix* provide additional details). Thus, most of the generated sequences adopt the DL fold on which they were optimized.

Of course, as shown by protein design studies, merely recovering active site residues is often insufficient to yield enzymatic function (52, 53). Other important factors include stabilizing the active site conformation so that the catalytically competent conformation is frequently populated and reproducing its electrostatic potential. Here, we consider the simplest requirement to generate minimal enzymatic function: the presence of appropriate residues with the correct chirality, geometry, and location. Given the plethora of sequences which nature could generate using foldamers or other means, likely the most catalytically active might survive. While we view this as a proof-of-principle study as what might have happened, we have predictions of stable sequences that should be examined experimentally (*SI Appendix*, LIST.stable_sequences).

A total of 472 of 550 (85.8%) active sites have at least one randomly generated sequence where all catalytic residues exactly match. If we further allow similar residues in the L amino acid active site positions, as assessed by a favorable BLOSUM 80 (54) score, then an additional 76 active sites are found. In total, 548 (99.6%) identical or similar active sites to the 550 native ones are found. Their E.C. numbers, number of catalytic residues, the number of sequences that either match the active site residues exactly or are similar, enzymatic function, and GO numbers (55, 56) are listed in *SI Appendix*, Table S1.

Fig. 5 shows the total number of sequences that match the different enzyme active sites. Obviously, the smaller the number of catalytic residues dictating the given E.C. number, NCAT, the easier it is to recover the active site. Thus, the number of sequences matching a given active site monotonically decreases as NCAT increases. For NCAT = 3, 229 of 235 E.C. numbers have exact sequence matches. There are a high of 9,017 sequences generated per active site to just 1. The other 6 three-residue active sites have similar sequence matches with 646 to 2 sequences per active site. For NCAT = 4, 170 distinct active sites have an exact residue match, with 919 to 1 sequences per active site independently generated. A total of 32 distinct active sites have similar residues with 723 to 1 sequences per active site. For NCAT = 5, 70 sites have an exact match with 38 to 1 sequences per site, while 35 sites have a match to similar sequences whose number ranges from 377 to 1. For NCAT = 6, 3 sites have a single exact sequence match.

Table 1 presents the subset of minimal bacterial enzymatic functions recovered in the demi-chiral protein sequence/structural library (31). Interestingly, many essential functions are found. These include enzymes associated with DNA repair, translation, protein processing, lipid metabolism, cofactor and nucleotide biosynthesis, and glycolysis. *SI Appendix*, Fig. S4 shows the subset of enzymes recovered (in red) associated with glycolysis and

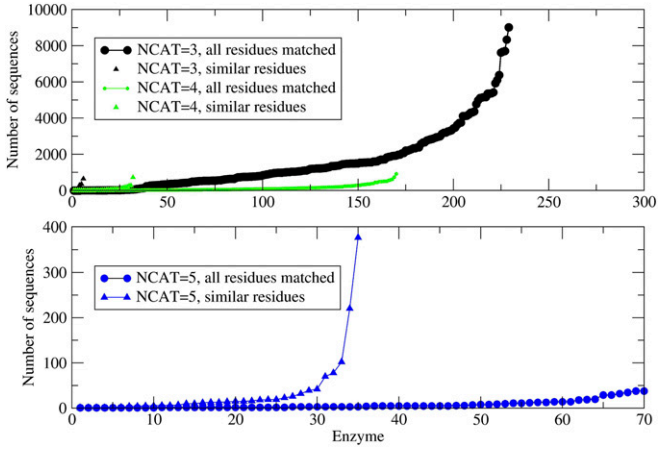

Fig. 5. For the number of catalytic residues, NCAT, ranging from 3 to 5, the number of sequences that match (triangles) or are similar (circles) to the active site residues of a particular enzymatic active site. The numbers on the abscissa are the label of the active site E.C. index for the given NCAT.

Table 1. List of enzymes recovered in the demi-chiral protein library that are members of the minimal bacterial gene set

Functional category	E.C. number	Biochemical function	No. of sequences	Exact match
DNA metabolism associated with the replication machinery	2.7.7.7	DNA-directed DNA polymerase	3	Yes
	6.5.1.2	DNA ligase	80	Yes
DNA repair				
	4.2.99.18	Class I DNA-(apurinic or apyrimidinic site) endonuclease	18	Yes
	3.2.2.23	DNA *N*-glycosylase	1,007	Yes
	3.2.2.21	DNA-3-methylguanine glycosylase	3	Yes
Translation: aminoacyl-t-RNA synthesis				
	6.1.1.1	Tyrosine-tRNA ligase	56	Yes
	6.1.1.6	Lysine-tRNA ligase	686	Yes
	6.1.1.10	Methionine-tRNA ligase	7	Yes
	6.1.1.11	Serine-tRNA ligase	19	Yes
	6.1.1.12	Aspartate-tRNA ligase	544	Yes
	6.1.1.18	Glutamine-tRNA ligase	966	Yes
	6.1.1.19	Arginine-tRNA ligase	1,901	Yes
	6.1.1.22	Asparagine-tRNA ligase	5,423	Yes
Ribosomal Function				
	2.1.1.48	tRNA (uracil-2′-*O*-)-methyltransferase	1	Yes
Protein processing				
	3.4.11.18	Aminopeptidase	82	Yes
	3.4.11.1	Aminopeptidase	1,363	Yes
Transport				
	2.7.1.69	D-glucosamine PTS permease	48	No
Glycolysis (8/10)				
	4.2.1.11	Enolase	2	Yes
	4.1.2.13	Fructose 1–6 biphosphate aldolase	3,412	Yes
	1.2.1.12	Glyceraldehyde 3-phosphase dehydrogenase	1,436	Yes
	5.4.2.1	Phosphoglycerate mutase	1,010	Yes
	1.1.1.27	L-lactate dehydrogenases	1,540	Yes
	2.7.1.11	6-Phosphofructokinase	10	Yes
	2.7.2.3	Phosphoglycerate kinase	2,054	Yes
	5.3.1.1	Triosephosphate isomerase	12	Yes
Pentose phosphate pathway				
	5.1.3.1	Ribulose-phosphate 3-epimerase	2	Yes
Lipid Metabolism				
	1.1.1.94	Glycerol-3-phosphate dehydrogenase	989	Yes
Biosynthesis of nucleotides				
	2.7.4.6	Nucleoside diphosphate kinase	1,709	Yes
	1.17.4.1	Ribonucleoside diphosphate reductase	38	Yes
	1.8.1.9	Thioredoxin-disulfide reductase	5,320	Yes
Biosynthesis of cofactors				
	2.7.7.3	Pantetheine-phosphate adenylyltransferase	341	Yes
	1.5.1.3	Dihydrofolate reductase	723	No
Methyl transferases				
	2.1.1.48	Adenine-N6-methyltransferase	1	Yes
	2.1.1.125	Histone-arginine *N*-methyltransferase	1,889	Yes
	2.1.1.107	Uroporphyrinogen-III C-methyltransferase	37	Yes
	2.1.1.63	Cysteine *S*-methyltransferase	196	Yes
	2.1.1.13	Methionine synthase	58	Yes
	2.1.1.20	Glycine methyl transferase	30	Yes
	2.1.1.43	Histone-lysine *N*-methyltransferase	1	Yes
	2.1.1.21	Methylamine-glutamate *N*-methyltransferase	46	No
	2.1.1.72	Site-specific DNA-methyltransferase (adenine-specific)	8	No

gluconeogenesis (57). Eight of 10 enzymes in the glycolysis pathway are found. While this pathway is not completely covered, it must be remembered that we have only explored enzymatic active sites located in a rather small set, 4,516, of demi-chiral protein structures and only those sites in the CSA active site library. One might expect the pathways to be fleshed out as additional demi-chiral structures are examined and the active site library size is increased.

Table 2 shows the distribution across the domains of life of the top 20 most frequently found enzymatic functions, with the top 40 listed in *SI Appendix*, Table S2. All contain 3 active site residues, and, consistent with Fig. 5, all 3 residues match those in

native protein active sites. Sixteen of 40 of these enzymes are found in Archaea, Bacteria, and Eukaryota, the 3 domains of life, with another 12 found in these domains as well as viruses. Presumably, these are ancient enzymes possibly present in the last common ancestor (32, 58, 59). Using BRENDA, we calculated that an average fraction of all enzymes found in all 3 domains of life is 0.26 (60). Combined with the results in Table 1, the calculated odds ratio of finding ancient enzymes in the top 40 E.C. numbers is 2.69, a significant enrichment over random. What is remarkable is that this result merely required the generation of compact demi-chiral protein structures with random, somewhat protein-like amino acid compositions and nothing more. At no time was selection for function done.

Table 3 shows the KEGG pathways in which the 550 generated enzymes participate, as well as their number per given pathway type (61). A very wide variety of metabolic processes are covered, including purine and pyruvate metabolism, sugar and amino acid metabolism, the citrate cycle, fatty acid biosynthesis, and lipid metabolism. As presented in additional detail in *SI Appendix*, Table S3, the average coverage of KEGG pathways by the "found" enzymes is 17.7%. This is not to say that all such enzymes (with marginal activity) were present in the putative demi-chiral protein soup, but, rather, these artificial demi-chiral proteins have the inherent capability of yielding a significant fraction of the biochemistry of life. Such biochemistry emerges from requirements that they must be present and have minimal stability and activity.

Conclusions

The lack of understanding of the origins of the breaking of demi-chirality found in the molecules of life on Earth is a long-standing problem, and models to date either focused on the RNA world hypothesis, which does not explain how RNA became chiral, or the use of chiral templates (e.g., chiral crystal surfaces). The alternative view due to Dyson conjectures that metabolism, likely from proteins, came first, followed by replication. But how did the ultimately homochiral proteins responsible for metabolism emerge from the short peptides that formed spontaneously and probably contained a mixture of D and L amino acids? The foldamer hypothesis suggests that such oligomers acted as templates to catalyze the synthesis of likely demi-chiral proteins. Other mechanisms such as molecular mutualism or the spontaneous peptide formation from aminonitriles might have been operative. By whatever means, we assume that, somehow, proteins, whose lengths range from 50 to 300 residues, were generated. This is the starting point for the present study.

Here, we explored the consequences if a fold library of initially demi-chiral proteins were generated without any selection for function but merely for the predicted thermodynamic stability of their compact structures. The results are both surprising and profound: the library of compact demi-chiral protein structures is predicted to be less stable than homochiral D or L proteins due to their reduced ability to form regular secondary structural elements because of lack of internal hydrogen bonding. Regular helical regions are shorter, and, while extended states exist, they are at best weakly hydrogen-bonded. On average, their predicted stability is 53% of native proteins, and they are relatively more stabilized by burial and pair interactions. This qualitative result is independent of the particular force field used, suggesting it is quite general and likely true. Improved hydrogen bonding drives selection toward more chiral systems, and, being more stable, they likely would have improved biochemical function.

Without any selection beyond predicted stability, the demi-chiral protein structural library displays a remarkable collection of native-like protein properties. It covers the space of all protein structures with significant structural (i.e., geometric) matches, including the global folds of early ribosomal proteins; which ones existed, of course, would have resulted from environmental circumstances. Essentially all contemporary ligand binding pockets have a structurally similar match to small molecule ligand binding pockets in demi-chiral proteins. Put another way, given the appropriate constellation of amino acids, demi-chiral proteins could generate the biochemical functions of contemporary proteins. But what about the presence of enzymes that engage in enantiospecific chemical reactions responsible for metabolism? Once again, as anticipated by protein design studies, even in a globally demi-chiral system, L amino acids are found in pockets whose RMSD lies within 1.0 Å of known active sites. Even for the small structural library considered here, with very restrictive geometric requirements and a limited active site library, the active sites corresponding to 550 active sites associated with 456 distinct E.C. numbers were found. Remarkably, when sequences of reasonable composition were randomly generated, again with no functional bias (we considered all contemporary amino acids to enable direct comparison to extant active sites), most active sites (~86%) have

Table 2. Distribution across the domains of life for the 40 most frequently found enzymatic functions in the demi-chiral protein library

Rank	E.C. number	No. of sequences	Enzymatic function	Domains of life
1	2.6.1.16	9,018	Glutamine-fructose-6-phosphate transaminase	Bacteria, Eukaryota
2	2.4.2.14	8,329	5-phospho-alpha-D-ribose 1-diphosphate	Bacteria/Eukaryota
3	4.1.99.3	7,707	Deoxyribodipyrimidine photo-lyase	Viruses and cellular organisms
4	1.18.6.1	7,670	Nitrogenase	Archaea, Bacterial, and Eukaryota
5	1.2.1.27	7,616	Methylmalonate-semialdehyde dehydrogenase (CoA-acylating)	Archaea, Bacterial, and Eukaryota
6	2.3.1.9	6,382	Acetyl-CoA C-acetyltransferase.	Archaea, Bacterial, and Eukaryota
7	3.4.22.15	6,104	Cysteine-type endopeptidase	Viruses and cellular organisms
8	4.1.1.8	5,930	Oxalyl-CoA decarboxylase	Bacteria, Eukaryota
9	4.2.1.71	5,420	Hydrolyase	Viruses and cellular organisms
10	1.8.1.9	5,360	Thioredoxin-disulfide reductase	Archaea, Bacterial, and Eukaryota
11	1.1.1.6	5,353	NAD-linked glycerol dehydrogenase	Bacteria, Eukaryota
12	1.2.2.2	5,369	Pyruvate dehydrogenases	Viruses and cellular organisms
13	3.1.1.4	5,158	Phospholipase A2	Viruses and cellular organisms
14	1.2.1.5	5,136	Aldehyde dehydrogenases	Archaea, Bacterial, and Eukaryota
15	4.1.3.18	5,114	(R)-citramalyl-CoA lyase	Viruses and cellular organisms
16	2.7.2.2	4,984	Carbamate kinases	Archaea, Bacterial, and Eukaryota
17	3.4.22.60	4,779	Cysteine-type endopeptidase	Tetrapoda
18	3.8.1.5	4,329	Haloalkane dehalogenases	Bacteria, Eukaryota
19	1.18.99.1	4,365	Ferredoxin hydrogenases	Viruses and cellular organisms
20	2.6.1.52	4,303	Phosphoserine transaminase	Archaea, Bacterial, and Eukaryota

Table 3. Summary of KEGG pathway types and number of matching enzymes found in the demi-chiral protein library

No. of Enzymes	Pathway type
246	Metabolic pathways
118	Biosynthesis of secondary metabolites
91	Microbial metabolism in diverse environments
27	Glycolysis/gluconeogenesis
21	Purine metabolism
20	Fructose and mannose metabolism
20	Carbon fixation in photosynthetic organisms
18	Pyruvate metabolism
18	Amino sugar and nucleotide sugar metabolism
17	Cysteine and methionine metabolism
16	Drug metabolism–other enzymes
16	Arginine and proline metabolism
16	Alanine, aspartate and glutamate metabolism
15	Pyrimidine metabolism
14	mTOR signaling pathway
14	Glyoxylate and dicarboxylate metabolism
14	Glycine, serine and threonine metabolism
13	Propanoate metabolism
12	alpha-Linolenic acid metabolism
12	Tryptophan metabolism
12	Pentose phosphate pathway
12	Methane metabolism
11	Starch and sucrose metabolism
11	Glutathione metabolism
11	Galactose metabolism
11	Citrate cycle (TCA cycle)
10	Valine, leucine and isoleucine degradation
10	Phenylalanine metabolism
10	PI3K-Akt signaling pathway
10	Glycerophospholipid metabolism
10	Aminoacyl-tRNA biosynthesis
9	Tyrosine metabolism
9	Phenylalanine, tyrosine, and tryptophan biosynthesis
9	Pentose and glucuronate interconversions
9	Nitrogen metabolism
9	Metabolism of xenobiotics by cytochrome P450
9	Lysine degradation
9	Folate biosynthesis
9	Fatty acid metabolism
9	Arachidonic acid metabolism
8	Glycerolipid metabolism
8	Carbon fixation pathways in prokaryotes
7	Steroid hormone biosynthesis
7	Penicillin and cephalosporin biosynthesis
7	Lysine biosynthesis
7	Fatty acid biosynthesis
7	Butanoate metabolism
7	Biotin metabolism
7	Benzoate degradation

exact sequence matches, and all but 2 active sites have matches if similar amino acids are also allowed. Importantly, the E.C. numbers with the largest number of generated sequences whose active site residues match native ones are highly enriched toward essential, ancient protein functions such as glycolysis, ribosomal function, translation, and DNA synthesis. We are not stating that all such functions were present at the origin of life, but rather that there is the inherent capacity to possess such functions, likely at a low level, if the relevant demi-chiral protein structures were present.

If proteins were randomly generated, by, say, the foldamer or aminonitrile hypotheses, such proteins could synthesize chiral molecules. Due to asymmetry in D:L amino acid composition in meteorites or even in a demi-chiral system, some proteins might possess an excess of D or L amino acids. These proteins would be more stable, and thus stability would drive selection toward more chiral systems. Perhaps a random fluctuation caused L-chirality to win. What then emerges is suggestive of a synthesis of the RNA and metabolism-first world ideas. These early proteins, while not as stable or functionally efficient as modern proteins, could engage in ancient metabolism, yielding lipids which could form vesicles as well as synthesizing chiral RNA. In other words, early metabolism could yield chiral RNA as one of its byproducts, which then eventually combined with the early universal, ribosomal proteins (also present in demi-chiral structures) to enable more efficient, more chiral protein synthesis. The present work suggests that t-RNA and DNA ligases are quite easy to generate in the demi-chiral protein library, but that DNA polymerases are harder to find as they contain more key active site residues; yet, they are there. This could generate a positive feedback loop, which explains how the breaking of chirality and emergence of metabolism and replication could have occurred quite close together in the primordial soup. One might imagine that the synthesized lipids formed vesicles, which then concentrated the relevant components for more efficient synthesis. The present study is, of course, theoretical and is a proof of principle. It is essential that these ideas be experimentally tested, but this study provides a well-defined road map as to how to do this.

Materials and Methods

Model Generation. The list of model proteins is provided in *SI Appendix*, LIST.models. The model generation process is described in detail in the *SI Appendix*, with an overview provided in *SI Appendix*, Fig. S1. The protein length distributions and local secondary structure biases of a given protein are taken from native structures. In practice, the resulting secondary structure in the folded protein often differs substantially from the initial secondary structure bias, and, as long as this is reasonable, it has little effect on the overall results. For all but 3 of 4,516 D:L proteins, the folded conformation has a TM-score <0.4 to the native protein that provided the local secondary structural bias. Next, a random pattern of L and D residues at the specified D:L ratio was generated for each protein. Each protein is represented in a main chain and side chain center of mass representation. We used the native amino acid geometry for the L chiral main-chain geometry and its mirror image for D chirality. Artificial polyleucine homopolymers are folded because polyleucine has the same compact global volume and small molecule ligand binding pocket volume as in native proteins (34). Structure models for D, L, and mixed L and D sequences were generated using a fragment-based ab initio method modified from the chunk-TASSER de novo protein structure prediction algorithm (45). First, a fragment library is generated by a modified SP^3 threading method (45, 62) using only secondary structure-dependent scores (i.e., no native amino acid sequence information was used) to select local protein fragments. Then, starting from random structures, ab initio chunk-TASSER randomly samples these local fragments to assemble compact global structures. A total of 1,000 models were generated for each protein. The structures were clustered using SPICKER (63). Next, SCWRL4 (64) was used to build full-atom polyleucine models. Finally, the DFIRE all-atom statistical potential (41) was employed to select the top polyleucine model for each protein.

Enzyme Active Sites. We employed a curated set of enzymes from the Catalytic Site Atlas (CSA) database (50). Each entry corresponds to a protein chain with an experimentally determined structure in the PDB (44), with manually annotated active sites obtained by literature mining. The CSA library contains 593 unique 4-E.C.-digit enzyme functions. We searched all selected polyleucine structures for amino acids with similar geometry and L amino acids as in the enzyme active sites in the CSA database. For each modeled compact structure, we first detect pockets using the geometry-based method LIGSITE (65), chosen because it is a very sensitive pocket detection algorithm. Here, we wanted a very rigorous definition of enzyme active site geometries. We then scanned these pockets against known active sites of the template library of enzymes with the pocket comparison program APoc (48). If the structure has an L-amino acid arrangement with a similar geometry as the active sites of a native enzyme whose RMSD from the known enzyme's active site is <1.0 Å, then we consider it a geometric hit. The list of enzymes along with the active site locations are provided in *SI Appendix*, LIST.enzymes.

Generating Protein Sequences for Analysis of Active Site Residues. For each polyleucine structure, a protein with a random composition and order of all 20 amino acid types was initially generated. Residue chirality is identical to that in the corresponding polyleucine structure. Twenty amino acids are used to allow comparison to contemporary active site residue types. To generate the initial protein sequence, rather than using the observed native amino acid composition, the amino acid composition was shuffled to reduce native-like amino acid composition biases. (We found that the results for the relative frequency of D:L sequences that match active sites are virtually the same if the average native amino composition is used.) A uniform random number assigns the amino acid type at each position. Then, as previously described (34, 36, 51), using a genetic algorithm, the resulting sequence order is optimized using secondary structure propensities (modified for mixed D and L systems), burial, and pair statistical potentials. Then, the predicted lowest-energy sequence is examined to see if it has the same active site residues as in the corresponding native protein's active site. For each of the 413 proteins whose structures contain active site geometries, 34,710,000 random sequences were generated.

Data Availability Statement. All data are provided in the *SI Appendix* and Datasets S1–S5.

ACKNOWLEDGMENTS. This work was supported in part by the Division of General Medical Sciences of the National Institutes Health (NIH Grant R35-118039).

1. N. Fujii, T. Saito, Homochirality and life. *Chem. Rec.* **4**, 267–278 (2004).
2. S. K. Kim, T. Ha, J. P. Schermann, Homochirality and the origin of life. Editorial of the PCCP themed issue. *Phys. Chem. Chem. Phys.* **13**, 804–805 (2011).
3. A. J. MacDermott *et al*., Homochirality as the signature of life: The SETH cigar. *Planet. Space Sci.* **44**, 1441–1446 (1996).
4. J. Podlech, New insight into the source of biomolecular homochirality: An extraterrestrial origin for molecules of life? *Angew. Chem. Int. Ed. Engl.* **38**, 477–478 (1999).
5. M. Wu, S. I. Walker, P. G. Higgs, Autocatalytic replication and homochirality in biopolymers: Is homochirality a requirement of life or a result of it? *Astrobiology* **12**, 818–829 (2012).
6. G. F. Joyce, The rise and fall of the RNA world. *New Biol.* **3**, 399–407 (1991).
7. G. F. Joyce, Building the RNA world. Ribozymes. *Curr. Biol.* **6**, 965–967 (1996).
8. G. F. Joyce, The antiquity of RNA-based evolution. *Nature* **418**, 214–221 (2002).
9. M. P. Robertson, G. F. Joyce, The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, a003608 (2012).
10. G. F. Joyce, J. W. Szostak, Protocells and RNA self-replication. *Cold Spring Harb. Perspect. Biol.* **10**, a034801 (2018).
11. F. J. Dyson, A model for the origin of life. *J. Mol. Evol.* **18**, 344–350 (1982).
12. K. Kvenvolden *et al*., Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature* **228**, 923–926 (1970).
13. J. R. Cronin, S. Pizzarello, Amino acids in meteorites. *Adv. Space Res.* **3**, 5–18 (1983).
14. T. N. Chiesl *et al*., Enhanced amine and amino acid analysis using Pacific Blue and the Mars Organic Analyzer microchip capillary electrophoresis system. *Anal. Chem.* **81**, 2537–2544 (2009).
15. J. E. Elsila *et al*., Meteoritic amino acids: Diversity in compositions reflects parent body histories. *ACS Cent. Sci.* **2**, 370–379 (2016).
16. A. Lupas, "At the origin of life: How did folded proteins evolve?" in *Research in Computational Molecular Biology*, M. Vingron, L. Wong, Eds. (Springer, New York, NY, 2008), vol. 4955, pp. 272.
17. V. Alva, A. N. Lupas, From ancestral peptides to designed proteins. *Curr. Opin. Struct. Biol.* **48**, 103–109 (2018).
18. E. Guseva, R. N. Zuckermann, K. A. Dill, Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7460–E7468 (2017).
19. K. A. Lanier, A. S. Petrov, L. D. Williams, The central symbiosis of molecular biology: Molecules in mutualism. *J. Mol. Evol.* **85**, 8–13 (2017).
20. P. Canavelli, S. Islam, M. W. Powner, Peptide ligation by chemoselective aminonitrile coupling in water. *Nature* **571**, 546–549 (2019).
21. C. Chothia, A. V. Finkelstein, The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039 (1990).
22. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
23. S. B. Pandit, J. Skolnick, Fr-TM-align: A new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **9**, 531 (2008).
24. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
25. Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, J. Skolnick, On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2605–2610 (2006).
26. J. Skolnick, H. Zhou, M. Brylinski, Further Evidence for the Likely Completeness of the Library of Solved Single Domain Protein Structures, Further evidence for the likely completeness of the library of solved single domain protein structures. *J. Phys. Chem. B* **116**, 6654–6664 (2012).
27. Z. Zhang, M. G. Grigorov, Similarity networks of protein binding sites. *Proteins* **62**, 470–478 (2006).
28. M. Gao, J. Skolnick, A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.* **9**, e1003302 (2013).
29. O. Khersonsky, C. Roodveldt, D. S. Tawfik, Enzyme promiscuity: Evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**, 498–508 (2006).
30. D. Davidi, L. M. Longo, J. Jabłońska, R. Milo, D. S. Tawfik, A bird's-eye view of enzyme evolution: Chemical, physicochemical, and physiological considerations. *Chem. Rev.* **118**, 8786–8797 (2018).
31. R. Gil, F. J. Silva, J. Pereto, A. Moya, Determination of the core of a minimal bacterial gene set. *Microbiol Mol. Biol. Rev.* **68**, 518–537 (2004).
32. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).
33. J. Skolnick, A. K. Arakaki, S. Y. Lee, M. Brylinski, The continuity of protein structure space is an intrinsic property of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 15690–15695 (2009).
34. J. Skolnick, M. Gao, Interplay of physics and evolution in the likely origin of protein biochemical function. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 9344–9349 (2013).
35. J. Skolnick, M. Gao, A. Roy, B. Srinivasan, H. Zhou, Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function. *Bioorg. Med. Chem. Lett.* **25**, 1163–1170 (2015).
36. J. Skolnick, M. Gao, H. Zhou, How special is the biochemical function of native proteins? *F1000 Res.* **5**, 207 (2016).
37. A. V. Korobeinikova, M. B. Garber, G. M. Gongadze, Ribosomal proteins: Structure, function, and evolution. *Biochemistry (Mosc.)* **77**, 562–574 (2012).
38. E. J. Milner-White, M. J. Russell, Functional capabilities of the earliest peptides and the emergence of life. *Genes (Basel)* **2**, 671–688 (2011).
39. H. Edwards, S. Abeln, C. M. Deane, Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput. Biol.* **9**, e1003325 (2013).
40. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
41. H. Zhou, Y. Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726 (2002).
42. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* **86** (suppl. 1), 7–15 (2018).
43. M. Brylinski, M. Gao, J. Skolnick, Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function. *Phys. Chem. Chem. Phys.* **13**, 17044–17055 (2011).
44. P. W. Rose *et al*., The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2017).
45. H. Zhou, J. Skolnick, Ab initio protein structure prediction using chunk-TASSER. *Biophys. J.* **93**, 1510–1518 (2007).
46. J. M. Harms *et al*., Translational regulation via L11: Molecular switches on the ribosome turned on and off by thiostrepton and micrococcin. *Mol. Cell* **30**, 26–38 (2008).
47. D. Perez-Fernandez *et al*., 4′-O-substitutions determine selectivity of aminoglycoside antibiotics. *Nat. Commun.* **5**, 3112 (2014).
48. M. Gao, J. Skolnick, APoc: Large-scale identification of similar protein pockets. *Bioinformatics* **29**, 597–604 (2013).
49. S. Tonddast-Navaei, B. Srinivasan, J. Skolnick, On the importance of composite protein multiple ligand interactions in protein pockets. *J. Comput. Chem.* **38**, 1252–1259 (2016).
50. N. Furnham *et al*., The catalytic site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485–D489 (2014).
51. J. Skolnick, M. Gao, H. Zhou, On the role of physics and evolution in dictating protein structure and function. *Isr. J. Chem.* **54**, 1176–1188 (2014).
52. O. Khersonsky *et al*., Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.* **407**, 391–412 (2011).
53. G. Kiss, N. Celebi-Olcum, R. Moretti, D. Baker, K. N. Houk, Computational enzyme design. *Angew. Chem. Int. Ed. Engl.* **52**, 5700–5725 (2013).
54. W. R. Pearson, Selecting the right similarity-scoring matrix. *Curr. Protoc. Bioinformatics* **43**, 3.5.1–3.5.9 (2013).
55. M. Ashburner *et al*.; The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
56. The Gene Ontology Consortium, The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
57. M. Kanehisa, Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.* **1611**, 135–145 (2017).
58. M. Y. Galperin, E. V. Koonin, Divergence and convergence in enzyme evolution. *J. Biol. Chem.* **287**, 21–28 (2012).
59. E. V. Koonin, Carl Woese's vision of cellular evolution and the domains of life. *RNA Biol.* **11**, 197–204 (2014).
60. L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).
61. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
62. H. Zhou, Y. Zhou, Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**, 321–328 (2005).
63. Y. Zhang, J. Skolnick, SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* **25**, 865–871 (2004).
64. G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, Jr, Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778–795 (2009).
65. B. Huang, M. Schroeder, LIGSITEcsc: Predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct. Biol.* **6**, 19 (2006).