Amino Acids (2015) 47:147–152
DOI 10.1007/s00726-014-1849-1

ORIGINAL ARTICLE

Structural protein reorganization and fold emergence investigated through amino acid sequence permutations

Giovanni Minervini · Alessandro Masiero · Emilio Potenza · Silvio C. E. Tosatto

Received: 8 September 2014 / Accepted: 29 September 2014 / Published online: 21 October 2014

Abstract Correlation between random amino acid sequences and protein folds suggests that proteins autonomously evolved the most stable folds, with stability and function evolving subsequently, suggesting the existence of common protein ancestors from which all modern proteins evolved. To test this hypothesis, we shuffled the sequences of 10 natural proteins and obtained 40 different and apparently unrelated folds. Our results suggest that shuffled sequences are sufficiently stable and may act as a basis to evolve functional proteins. The common secondary structure of modern proteins is well represented by a small set of permuted sequences, which also show the emergence of intrinsic disorder and aggregation-prone stretches of the polypeptide chain.

Keywords Protein fold · Evolution · Origin of life · Ab initio · Structure prediction · Intrinsic protein disorder

Introduction

Proteins are considered as the building blocks of life owing to their stability and functional plasticity. At the molecular level, proteins are heteropolymers of 20 different amino acids. Simple dipeptides such as Gly-Gly or Ser-His were reported to catalyze both proteolytic and peptide bond formation (Li et al. 2000; Plankensteiner et al. 2002). It is conceivable that small catalytic proto-enzymes contributed to the evolution of longer copolymers, triggering the race for the most stable polypeptide chain. It is also conceivable that a stable chain incorporated the same catalytic motifs, perhaps yielding a first complex enzyme (e.g., a catalytic center coupled with a stable scaffold). Over time proteins evolved different and complex functions such as enzymatic catalysis, structural, and signaling functions. Current theories regarding the origin of life suggest that amino acid chains co-evolved with, or immediately after, the emergence of autocatalytic RNA (Zhang and Cech 1997; Johnston et al. 2001). On the basis of these theories, primordial proteins, also known as proteinoids, promoted the nucleic acid onset, allowing their stabilization and conferring the ability to adapt to environmental changes (Berger 2003). Ideally, primordial RNA–polypeptide association could have enlarged the spectrum of different catalyses allowing the formation of more complex reactions. Notably, in modern organisms peptide bond formation is driven by the ribosome, a highly efficient RNA–protein complex. Considering a prebiotic scenario where rigid environmental conditions coupled with an unstable chemical surrounding promoted protein degradation, a stable structure could be considered as a favorable requirement for protein evolution. Indeed, since the appearance of proteins, only few folds became fixed, perhaps as a result of thermodynamic (or biological *ante litteram*?) selection of the most stable three-dimensional structures. Correlations between random amino acid sequences and protein folds were investigated by Weiss et al. (2000), suggesting the general idea that proteins autonomously evolved the most stable folds only after sequence evolution "refined" the surviving folds in order to acquire stability and function (Minervini et al. 2009; Lavelle and Pearson 2010). Indeed, modern proteins are characterized by a large number of different sequences not balanced by the same number of

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1849-1) contains supplementary material, which is available to authorized users.

G. Minervini · A. Masiero · E. Potenza · S. C. E. Tosatto (✉)
Department of Biomedical Sciences, University of Padova,
Viale G. Colombo 3, 35131, Padua, Italy
e-mail: silvio.tosatto@unipd.it

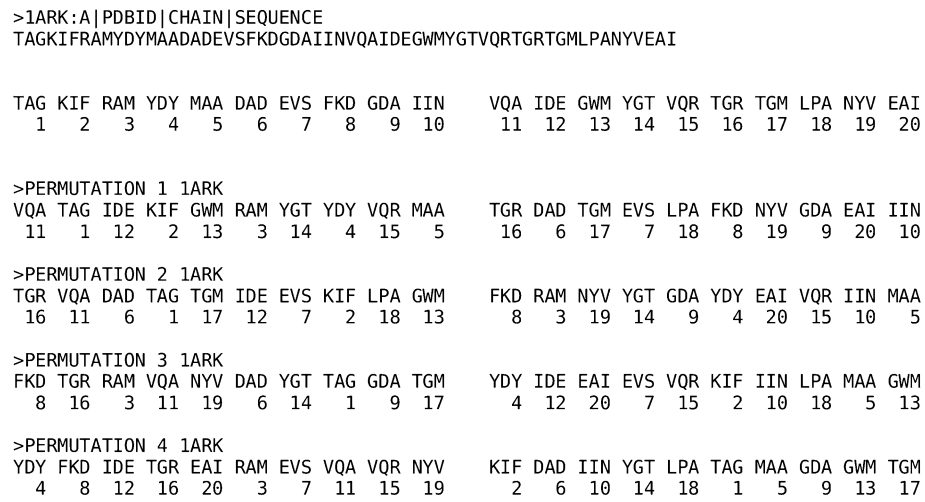

Fig. 1 Shuffling schema. Each sequence was initially split into 20 blocks of three amino acids long and then divided in two groups termed A and B. The first block of group B was placed at the beginning of the new permuted sequence and the first block of A was then placed second. The same was applied for the other blocks until a new permuted sequence was obtained

different folds (Pearl et al. 2005). Unexpectedly, structural analysis in silico on random proteins suggested that polypeptide chains may fold and adopt conformations comparable to natural proteins (De Lucrezia et al. 2012). Proteinoid theory assumes that simple enzymatic function evolved spontaneously from simple and not biologically generated polypeptides (Turian 1999). On the other hand, it is credible that the first proteins were generated through ligation of small and stable "amino acid modules" operated by proto-enzymes characterized by both low specificity and efficiency. In this scenario, small proto-proteins should be responsible for all resulting modern folds. In this work, we investigated the occurrence of different folds using an in silico combinatorial approach. We shuffled the sequences of 10 natural proteins and obtained 40 unrelated different folds. A large spectrum of different organizations were observed, including all-α, β-barrel, and disorder. Finally, our results suggest that abiotic complex fold emergence may have resulted from stochastic sequence rearrangement.

Methods

Permutations were manually performed starting from 10 different sequences in Fasta format of the same number of natural proteins selected from the Protein Data Bank (PDB codes: 1ARK, 1HX2, 1KV0, 1PPT, 1TCP, 1UOY, 1ZFI, 2BHI, 2CDX, 2KJF) (Berman et al. 2007). In order to minimize both the error and computational demands associated with the ab initio methodology, we selected proteins with a maximum length of 66 amino acids. Each sequence was initially split into 20 blocks of three amino acids long and then divided into two halves, termed A and B. The first block of group B was placed at the beginning of the new permuted sequence and the first block of A was then placed second. The same was applied for the other blocks until a new permuted sequence was obtained. A schematic explanation of the permutation schema is shown in Fig. 1. The three-dimensional model structures of both natural and permuted proteins were predicted using Rosetta (Rohl et al. 2004), a piece of ab initio protein structure prediction software based on the assumption that local interactions bias the conformation of sequence fragments in a polypeptide chain, while global interactions determine the three-dimensional structure with minimal energy (Rohl et al. 2004). For each sequence, 25,000 decoys were generated and clustered using the integrated clustering module. Only the best ranked model proposed for each sequence was considered. The predicted three-dimensional structures of native sequences were calculated and used to test the setup applied in the ab initio protocol. A comparison between experimental and predicted structures is shown in Supplementary Fig. 1. The overall fold stability was studied by performing 4 ns of molecular dynamics (MD) simulation with Gromacs (Hess et al. 2008) using the CHARMM 27 force field (MacKerell et al. 1998). At 4 ns per system the overall simulation time for all predictions was 200 ns. The total energy of the protein in solution was calculated with BLUUES (Walsh et al. 2012b). Disorder was predicted using ESpritz (Walsh et al. 2012a) and both secondary structure and aggregation propensity with PASTA 2.0 (Walsh et al. 2014).

Results

The aim of the present work was to elucidate the way evolution generated the variability observed among existing protein folds. In this sense, we hypothesized a credible scenario where several different sequences arose from a common mixture of short prebiotic polypeptide chains (e.g., a primordial broth). Starting from an existing protein, we split the sequence into smaller segments of three amino

acids each, then reassembled the blocks to obtain four new shuffled sequences. The resulting sequences conserved the same overall amino acid composition as the natural one, but different internal organization, as shown in Fig. 1.

Fold analysis

The propensity of shuffled proteins to generate stable folds was measured by calculating a three-dimensional structure for each permuted sequence with Rosetta (Rohl et al. 2004). Interpretations of the results described here are heavily related to the validity of structures predicted using the ab initio method. In a number of cases, Rosetta was shown to perform fairly well and even produced near-atomic resolution structures (Bradley et al. 2005; Das and Baker 2008). Our results for the natural proteins used in this research confirm that the predictions are in most cases accurate in terms of overall fold, secondary structure content, and topology (Supplementary Fig. 1). The same protocol was also applied to the shuffled sequences. The structures obtained show that pseudo-proteins tend to assume a well-ordered three-dimensional structure, with almost all predictions promoting a compact fold (Fig. 2). Notably, in several tests the structure obtained at permutation four resulted in a completely different secondary structure content with respect to the starting structure. This finding reinforces the idea that the propensity to evolve different folds is an emergent property of amino acid chains and clearly showed that a sequence can freely evolve different three-dimensional shapes when not constrained by the requirements of biological evolution (De Lucrezia et al. 2012). An illustration of this evidence may be given by the permutation of carnocyclin (PDB code 2KJF) (Fig. 2). The native protein folds in an all-α structure and is known to be very stable, with a fold shared with other proteins of the bacteriocin family (Martin–Visscher et al. 2009). Owing to its important biological function, the general fold of this protein family is largely conserved between bacteria. When permuted, the protein gave a wide spectrum of different organizations. Indeed, its permutation generates a β-barrel-like, plain β-sheet, α–β mixed, and finally an all-α structure, at permutation steps 1, 2, 3, and 4, respectively. Similar permutation behavior was obtained for a trypsin inhibitor (PDB 1HX2) and avian pancreatic polypeptide (PDB 1PPT). More generally, in all simulations we observed an alternation between several folds with the protein structure suddenly changing between permutations. The result confirms in part what was previously reported by Luisi and co-workers (Chiarabelli et al. 2006), i.e., the identification of folded proteins emerging from a random library. On the other hand, variability may suggest that different folds can spontaneously derive from a small sequence subset as obtained from only four permutation events.

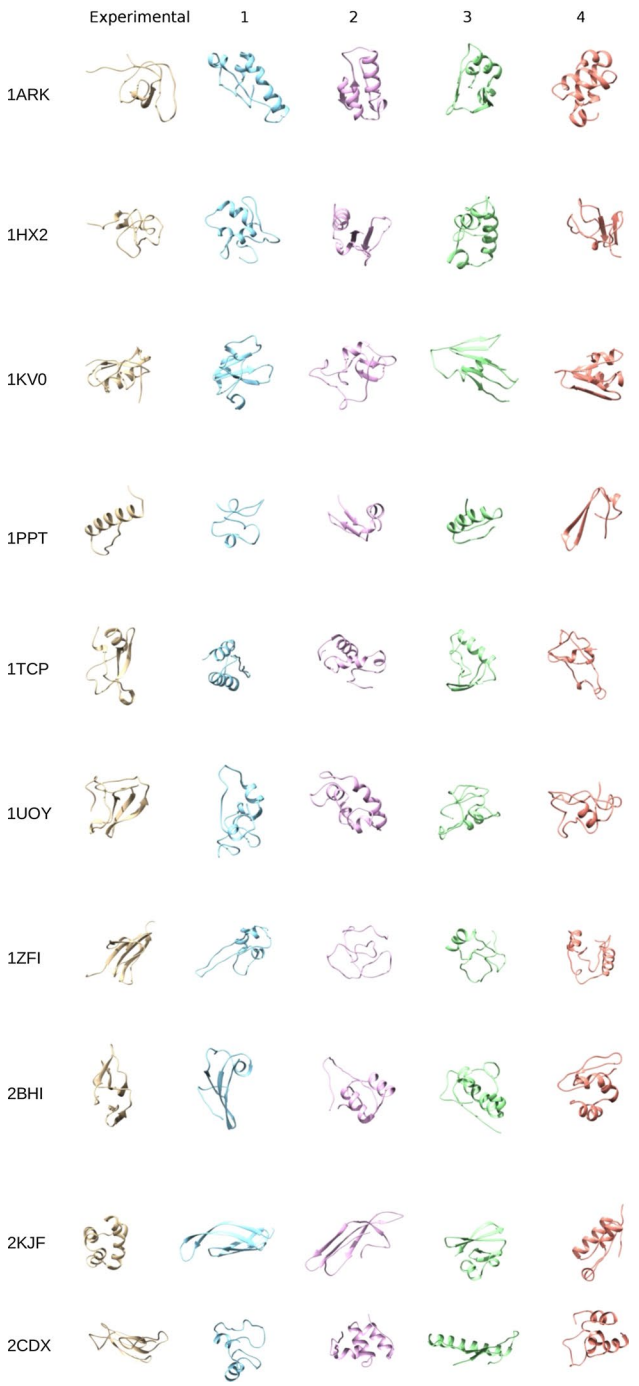

Fig. 2 Prediction of permuted proteins. Comparison between experimental structure and predicted structures after permutation. *Brown* represents the experimental structure; *blue*, *purple*, *green*, and *orange* represent permutations 1, 2, 3, and 4, respectively. The corresponding PDB code is reported on the *left*

Stability analysis

Every predicted protein, native or shuffled, went through a short (4 ns) MD simulation with the aim of evaluating the stability of the generated folds in aqueous solution.

Counting every generated sequence and prediction, 50 MD simulations were carried out and evaluated through both visual inspection and root mean square deviation (RMSD) plots. This helped direct visual detection of the most widely fluctuating systems. For the majority of the targets this revealed no wide nor fast misfolding behavior (Supplementary Fig. 2). The plots of the shuffled predictions in some cases show slightly higher values than the predictions, which is probably due to structural relaxation and energy minimization. This behavior is also confirmed by the total BLUUES (Walsh et al. 2012b) energy (see Fig. 3). A clear example can be seen in carnocyclin (PDB code 2KJF), where RMSD variation of the shuffled proteins reaches 500 % the native prediction, the highest overall value. This can be partially explained by the non-shuffled prediction RMSD plot, which has a rigid overall structure. This protein also shows the highest variability among different folds. However, no misfolding was observed during the simulations. Although the simulation time for each model was kept short because of the large number of systems to test, no misfolding events were observed and in almost all cases RMSD values were lower than 0.4 nm. Considering this result we can assume that permuted polypeptide chains show a structural stability comparable to the natural proteins.

Homology search and disorder analysis

In order to evaluate the effect of sequence similarity on the permuted sequences, we performed a similarity search with Blast (Altschul et al. 1997). The native sequences were all identified, but the search for permuted sequences gave no significant results. In other words, this finding confirmed that the tested sequences should be considered non-natural. Indeed, we were confident that structural predictions were not (or minimally) influenced by homology with other known proteins. The permutation of leech carboxypeptidase inhibitor (PDB 1ZFI) presents another peculiar result, showing a disordered fold for permutation 2 (Fig. 2). Similar results were also obtained for both avian pancreatic polypeptide (PDB 1PPT) and the so-called bubble protein (PDB 1UOY) at permutation 1. Intrinsically disordered proteins are biological entities showing important biological activities (i.e., protein–protein interaction) characterized by existing in a constitutively unfolded state (Iakoucheva et al. 2002). As Rosetta was not originally developed to predict disorder, we tested the sequences with ESpritz (Walsh et al. 2012a) to confirm this finding (see Fig. 3). The analysis revealed that only few pseudoproteins tend to assume a disordered state as the main secondary structure organization. The average disorder content (17.1 %) is slightly higher than the wild-type one (16 %), with the notable exception of PDB code 1PPT. A similar but stronger trend is also observed for aggregation propensities measured with PASTA 2.0 (Walsh et al. 2014), where the overall percentage of aggregating residues rises from 11.45 to 20.16 %, suggesting a negative selection for aggregating sequences. Although consistent with structure prediction, this finding should be considered only as preliminary, as low sequence similarity may induce the disorder and aggregation predictors to produce false positive output. Considering all results in a prebiotic context, we can speculate that the major structural organization observed in modern natural proteins can be evolved from a simple rearrangement of the same building blocks.

Discussion

Shared fold and sequence similarity are common traits of modern proteins as shown in, e.g., the SCOP database (Murzin et al. 1995). Current accepted theories explain this finding assuming a hypothetical common ancestor from which all modern organisms derived (Steel and Penny 2010; Theobald 2010). In other words, all current folds could have been fixed when life emerged on Earth and survived until today owing to their intrinsic stability. In this work, we explored one of the possible solutions adopted by nature to generate different folds from prebiotic-compatible building blocks. We shuffled the sequences of natural proteins in order to obtain new permutated proteins. In 2002, spontaneous formation of a chain of 10–15 amino acids long in rigid prebiotic conditions was reported (Commeyras et al. 2002). The authors suggested that evolution of large proteins is compatible with a fickle environment, at least chemically speaking. In this sense, we hypothesized a scenario where different proto-proteins arose from a mixture of short prebiotic polypeptides. In previous work, Kauffman (2003) argued that the ability to act autonomously in an environment is a fundamental characteristic of life. In other words, both diverse and stable folds provide the stability needed for effective biological evolution. In 2006, Luisi and co-workers reported that random small polypeptides spontaneously tend to assume both a compact and stable conformation (Chiarabelli et al. 2006). It is useful to recall that modern proteins share only a limited number of folds, generally combined to form multi-domain architectures (Pearl et al. 2005). One can argue that this is due to the biological evolution, with modern folds representing the best solution found by the nature to solve a specific problem. Our opinion is that the modern folds may result from biological evolution operating in the form of continuous finishing of a finite set of randomly generated proto-folds. We demonstrated that shuffled sequences may act as a basis to evolve functional proteins, with the common secondary structure of modern proteins well represented by a small set of permuted sequences. Although meaningful, the results presented here

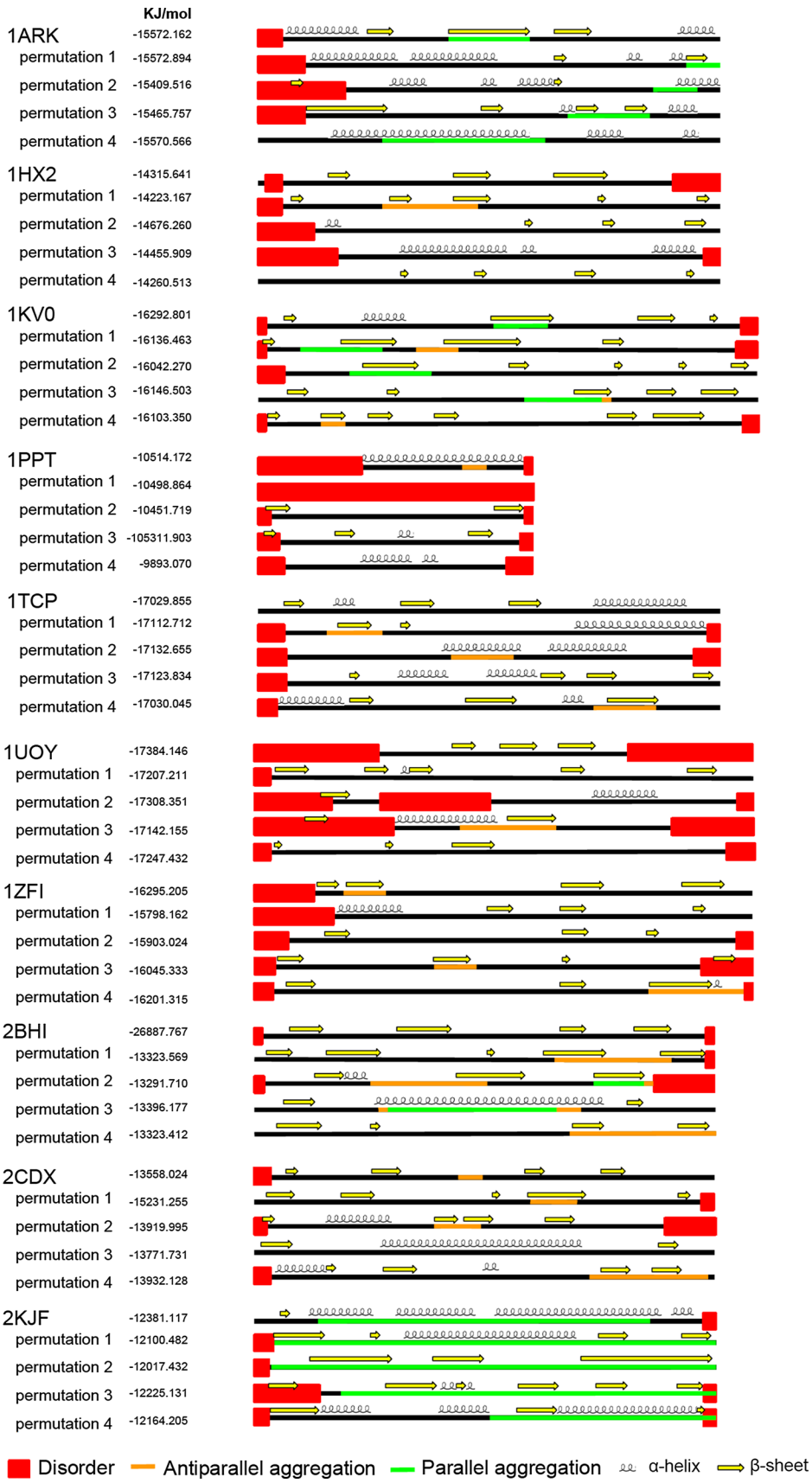

Fig. 3 Sequence features and energy of the permuted proteins. A sequence diagram is shown for each experimental protein and its four permutations, highlighting predicted secondary structure (*spirals* for helices and *arrows* for strands), disorder (*thick red line*), and aggregation propensity (*narrow yellow* and *green lines*) according to the symbols explained at the bottom of the figure. The total BLUUES energy of the protein structure is shown after the sequence identifier (color figure online)

are theoretical and their reliability is related to the validity of the underlying in silico predictions. Previous work suggested that folding is an innate property of amino acid chains (Chiarabelli et al. 2006; Minervini et al. 2009; LaBean et al. 2011). Nevertheless, we found several sequences apparently promoting disorder and characterized by low complexity. During the last decade, disorder assumed relevance in protein science with a large number of intrinsically unstructured proteins discovered to be involved in fundamental cellular processes such as protein–protein interactions, scaffolding, signaling, and transcription (Iakoucheva et al. 2002). Our research suggested that both structured and intrinsically unfolded proteins can evolve by means of simple amino acid recombination. This finding, if experimentally validated, will perhaps provide a better understanding of the driving forces favoring the emergence of Life.

Acknowledgments The authors are grateful to members of the BioComputing UP lab for insightful discussions. GM is an AIRC research fellow.

Conflict of interest The authors declare that they have no conflict of interest.

References

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Berger G (2003) Deterministic hypotheses on the origin of life and of its reproduction. Med Hypotheses 61:586–592. doi:10.1016/S0306-9877(03)00237-8

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303. doi:10.1093/nar/gkl971

Bradley P, Malmström L, Qian B et al (2005) Free modeling with Rosetta in CASP6. Proteins 61(Suppl 7):128–134. doi:10.1002/prot.20729

Chiarabelli C, Vrijbloed JW, De Lucrezia D et al (2006) Investigation of de novo totally random biosequences, Part II: on the folding frequency in a totally random library of de novo proteins obtained by phage display. Chem Biodivers 3:840–859. doi:10.1002/cbdv.200690088

Commeyras A, Collet H, Boiteau L et al (2002) Prebiotic synthesis of sequential peptides on the Hadean beach by a molecular engine working with nitrogen oxides as energy sources. Polym Int 51:661–665. doi:10.1002/pi.1027

Das R, Baker D (2008) Macromolecular modeling with rosetta. Annu Rev Biochem 77:363–382. doi:10.1146/annurev.biochem.77.062906.171838

De Lucrezia D, Slanzi D, Poli I et al (2012) Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. PLoS One 7:e36634. doi:10.1371/journal.pone.0036634

Hess B, Kutzner C, van der Spoel, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447

Iakoucheva LM, Brown CJ, Lawson JD et al (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323:573–584. doi:10.1016/S0022-2836(02)00969-5

Johnston WK, Unrau PJ, Lawrence MS et al (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 292:1319–1325. doi:10.1126/science.1060786

Kauffman S (2003) Molecular autonomous agents. Philos Trans R Soc Lond Ser Math Phys Eng Sci 361:1089–1099. doi:10.1098/rsta.2003.1186

LaBean TH, Butt TR, Kauffman SA, Schultes EA (2011) Protein folding absent selection. Genes 2:608–626. doi:10.3390/genes2030608

Lavelle DT, Pearson WR (2010) Globally, unrelated protein sequences appear random. Bioinformatics 26:310–318. doi:10.1093/bioinformatics/btp660

Li Y, Zhao Y, Hatfield S et al (2000) Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. Bioorg Med Chem 8:2675–2680

MacKerell Bashford D, Bellott et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616. doi:10.1021/jp973084f

Martin–Visscher LA, Gong X, Duszyk M, Vederas JC (2009) The three-dimensional structure of carnocyclin A reveals that many circular bacteriocins share a common structural motif. J Biol Chem 284:28674–28681. doi:10.1074/jbc.M109.036459

Minervini G, Evangelista G, Villanova L et al (2009) Massive non-natural proteins structure prediction using grid technologies. BMC Bioinformatics 10(Suppl 6):S22. doi:10.1186/1471-2105-10-S6-S22

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540. doi:10.1006/jmbi.1995.0159

Pearl F, Todd A, Sillitoe I et al (2005) The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33:D247–D251. doi:10.1093/nar/gki024

Plankensteiner K, Righi A, Rode BM (2002) Glycine and diglycine as possible catalytic factors in the prebiotic evolution of peptides. Orig Life Evol Biosphere J Int Soc Study Orig Life 32:225–236

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93. doi:10.1016/S0076-6879(04)83004-0

Steel M, Penny D (2010) Origins of life: common ancestry put to the test. Nature 465:168–169. doi:10.1038/465168a

Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature 465:219–222. doi:10.1038/nature09014

Turian G (1999) Origin of life. II. from prebiotic replicators to protocells. Arch Sci Compte Rendu Seances Soc 52:101–109

Walsh I, Martin AJM, Di Domenico T, Tosatto SCE (2012a) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28:503–509. doi:10.1093/bioinformatics/btr682

Walsh I, Minervini G, Corazza A et al (2012b) Bluues server: electrostatic properties of wild-type and mutated protein structures. Bioinformatics 28:2189–2190. doi:10.1093/bioinformatics/bts343

Walsh I, Seno F, Tosatto SCE, Trovato A (2014) PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res 42:W301–W307. doi:10.1093/nar/gku399

Weiss O, Jiménez-Montaño MA, Herzel H (2000) Information content of protein sequences. J Theor Biol 206:379–386. doi:10.1006/jtbi.2000.2138

Zhang B, Cech TR (1997) Peptide bond formation by in vitro selected ribozymes. Nature 390:96–100. doi:10.1038/36375