JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
2020, VOL. 38, NO. 15, 4579–4583
https://doi.org/10.1080/07391102.2019.1682051

Taylor & Francis
Taylor & Francis Group

EXPRESS COMMUNICATION

What limits the primary sequence space of natural proteins?

Aditya Mittal[a], Anandkumar Madhavjibhai Changani[a] and Sakshi Taparia[b]

[a]Kusuma School of Biological Sciences, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India; [b]Department of Mathematics, Bachelors Program in Mathematics & Computing, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

Number of naturally occurring primary sequences of proteins is an infinitesimally small subset of the possible number of primary sequences that can be synthesized using 20 amino acids. Prevailing views ascribe this to slow and incremental mutational/selection evolutionary mechanisms. However, considering the large number of avenues available in form of diversity of emerging/evolving and/or disappearing living systems for exploring the primary sequence space over the evolutionary time scale of ~3.5 billion years, this remains a conjecture. Therefore, to investigate primary sequence space limitations, we carried out a systematic study for finding primary sequences absent in nature. We report the discovery of the smallest peptide sequence "Cysteine-Glutamine-Tryptophan-Tryptophan" that is not found in over half-a-million curated protein sequences in the Uniprot (Swiss-Prot) database. Additionally, we report a library of 83605 pentapeptides that are not found in any of the known protein sequences. Compositional analyses of these absent primary sequences yield a remarkably strong power relationship between the percentage occurrence of individual amino acids in all known protein sequences and their respective frequency of occurrence in the absent peptides, regardless of their specific position in the sequences. If random evolutionary mechanisms were responsible for limitations to the primary sequence space, then one would not expect any relationship between compositions of available and absent primary sequences. Thus, we conclusively show that stoichiometric constraints on amino acids limit the primary sequence space of proteins in nature. We discuss the possibly profound implications of our findings in both evolutionary and synthetic biology.

ARTICLE HISTORY
Received 23 August 2019
Accepted 11 October 2019

KEYWORDS
Primary sequence; protein; de novo design; peptide

Introduction

There has been a recent burst of exciting research in *de novo* protein design. From creation of small peptides with specific structural and functional features (Bhardwaj et al., 2016; Chevalier et al., 2017) to engineering of larger assemblies with specific functionalities (Boyken et al., 2019; Dou et al., 2018; Koepnick et al., 2019; Langan et al., 2019; Marcos et al., 2018; Ng et al., 2019; Shen et al., 2018), *de novo* protein design is pushing the boundaries of the evolutionarily available protein space. In large part, this was pre-empted by a review (Huang, Boyken, & Baker, 2016) in which the scope of naturally unexplored protein primary sequence space was highlighted – this was numerically in contrast to Levinthal's observations on the large conformational space that cannot be explored by primary sequences (Levinthal, 1968; 1969). On one hand, Levinthal's paradox (Levinthal, 1968; 1969) captures the importance of structural constraints in protein folding– complete conformational space available to a protein sequence cannot be sampled to arrive at a single functional conformation within time constraints of life. On the other hand, it was surmised that the primary sequence space of proteins is actually not constrained (Huang et al., 2016) – it is conjectured that the slow and incremental evolutionary mechanisms have allowed only a limited exploration of the primary sequence by living systems using 20 amino acids. However, is the natural proteins' primary sequence space really beyond any constraints? Till date there is no systematic and direct study on exploring limitations of the primary sequence space of natural proteins, if any exist. The only indirect investigations have been dominated by deriving propensities/statistical potentials for amino acids (and/or their groups) based on their presence in specific structural features (Chothia, 1992; Karplus & Kuriyan, 2005; Pauling, Corey, & Branson, 1951; Rose, Fleming, Banavar, & Maritan, 2006). The first clues on possible compositional constraints in primary sequences of naturally occurring folded proteins were provided by the discovery of "stoichiometric margins of life" – referring to lower than statistically expected standard deviations of mean occurrences of amino acids in sequences of naturally occurring folded proteins (Mezei, 2011; Mittal, Jayaram, Shenoy, & Bawa, 2010; Mittal & Jayaram, 2011a; 2011b). However those compositional constraints were derived from structural data, without considering the order of amino acids occurring in primary sequences. Therefore, they also do not address possible limitations to the primary sequence space in natural proteins, if any exist. In complete contrast to the above, systematic investigations for several

CONTACT Aditya Mittal amittal@bioschool.iitd.ac.in Kusuma School of Biological Sciences, Indian Institute of Technology Delhi (IIT Delhi), Hauz Khas, New Delhi 110016, India

decades have uncovered conformational limitations for natural proteins thereby moving towards resolving of the Levinthal's paradox. While Ramachandran's plot (Ramachandran, Ramakrishnan, & Sasisekharan, 1963) provided steric constraints applicable to amino acids in folded protein structures, Anfinsen's experiments (Anfinsen, 1973; Anfinsen, Haber, Sela, & White, 1961) provided insights into constraints on the structural space of folded proteins via energy-driven conformational memory. Subsequently, studies on the protein folding problem have focused, over the years, on structure prediction from a given primary sequence based on searching for lowest possible energy configurations (Baldwin, 1995; Dill & Chan, 1997; Huang et al., 2016; Wolynes, Onuchic, & Thirumalai, 1995).

In view of the above, we carried out a systematic investigation on exploring possible constraints to the primary sequence space of natural proteins. First we computationally synthesized all possible peptides up to pentapeptides using the 20 amino acids. That gave us over three million peptides (3368400). Then we counted the number of times each of these peptides occur in more than half-a-million (560459) "manually annotated and reviewed" (curated) primary sequences from Uniprot: Swiss-Prot (UniProt Consortium, 2019), using every possible reading frame of the primary sequence depending on the length of the peptide being searched. We report the remarkable discovery of one (01) tetrapeptide and over eighty thousand pentapeptides that do not occur in these known primary sequences. We also report that the total number of times amino acids occur in the absent pentaptides has a very strong mathematical relationship with the percentage occurrence of the respective amino acids in all known primary sequences. This conclusively demonstrates that the primary sequence space of natural proteins is largely limited by relative abundance of the available amino acids. Finally, the discovery of library of small peptides (1 tetra- and 83605 penta- peptides) not found in nature is expected to provide strong support to the efforts of *de novo* protein design beyond the natural collection.

Methods

Complete sequence data was downloaded from Uniprot (Swiss-prot) on 28th July 2019 as per instructions provided for offline analyses. Results presented are from this dataset. The list of all computationally synthesized peptides (400 dipeptides, 8000 tripeptides, 160000 tetrapeptides and 3200000 pentapeptides) and their respective number of occurrences in all curated sequences in Uniprot (Swiss-Prot) are provided in ".csv" format at the following link: "http://web.iitd.ac.in/~amittal/Data_Mittal_etal_JBSD.html".

The percentage occurrence of amino acids in all known curated sequences was obtained directly from "Amino acid distribution statistics" in the "Statistics" link – "UniProtKB/Swiss-Prot UniProt release 2019_07 Jul-31, 2019". An earlier download of complete sequence data on 15th February 2019 was done to develop and test the analytical codes. Negligible differences (almost nil) in the results were obtained from the two datasets. Coding was done in Python for counting the number of occurrences of peptides. Independent coding was done in Java to confirm accuracy of the results.

Results and discussion

Computational synthesis of peptides and searching for their presence

We first downloaded all the curated sequences available in Uniprot (Swiss-Prot); total sequences = 560459. Next, using the 20 standard natural amino acids, we computationally synthesized libraries of dipeptides (total = 20x20 = 400), tripeptides (total = 20x20x20 = 8000), tetrapeptides (total = 20x20x20x20 = 160000) and pentapeptides (total = 20x20x20x20x20 = 3200000). Then, by reading each sequence from amino-to-carboxy terminals, we counted number of times each peptide occurred all of the sequences – libraries of the synthesized peptides and their respective number of occurrences can be downloaded from "http://web.iitd.ac.in/~amittal/Data_Mittal_etal_JBSD.html". For the purpose of counting presence of a peptide, number of reading frames for each sequence was equal to the peptide size. Thus, to count number of times a given dipeptide occurs in a given sequence, two reading frames starting from the first two amino acids respectively were utilized, and so on as shown in Figure 1a. The primary goal was to identify those peptides that were absent from the Uniprot (Swiss-Prot) database. Therefore, we identified how many, and which, peptides did not occur even once in the sequence data. While all 400 dipeptides and 8000 tripeptides were found to be present multiple times, exactly one (01) tetrapeptide and 83605 pentapeptides were absent in the sequences, regardless of reading frames. The one tetrapeptide that was absent is "CQWW" (Cysteine-Glutamine-Tryptophan-Tryptophan). The list of absent pentapeptides, in ".csv" format, is available for downloading at "http://web.iitd.ac.in/~amittal/Data_Mittal_etal_JBSD.html".

Stoichiometric constraints on amino acids limit the natural primary sequence space

The next step was detailed compositional analyses of the absent peptides along with investigation of possible physicochemical reasons for their absence. From the compiled list of the absent pentapeptides, we first counted number of times each of the 20 residues appeared at positions 1 to 5. To our surprise, we found that the number of occurrences for any given residue in any of the five positions was similar, as shown in Figure 1b. While all 20 amino acids have different number of occurrences in the absent pentapeptides, each amino acid occurs almost equal number of times in each of the five positions of the absent pentapeptides. This implies that the probability of occurrence of a given amino acid was the same in the absent peptide regardless of its position. In order to understand this apparently strange observation, we plotted the average occurrence of each amino acid in the absent pentapeptides vs. percentage occurrence of that amino acid in the Uniprot (Swiss-Prot) database. Figure 1c shows that there is a remarkably strong mathematical power relationship between

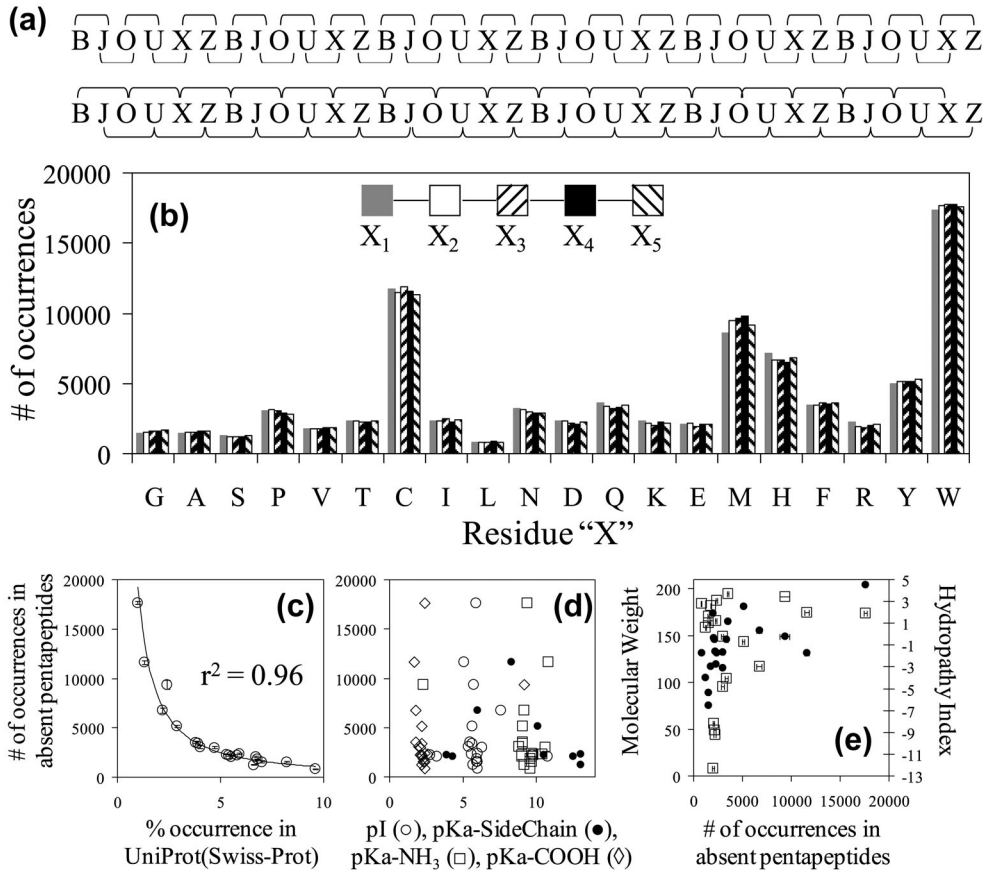

Figure 1. Counting peptides in primary sequences and compositional analyses of absent pentapeptides. **(a)** Examples are shown for dipeptides and tripeptides. Non-standard amino acid letters, i.e. BJOUXZ, were not counted for compiling the number of times a given peptide occurs in a given sequence – the total occurrences of non-standard amino acid letters was negligible (≪0.5%). **(b)** Total number of occurrences of each amino acid residue at positions 1 to 5 (X_1 to X_5) in 83605 pentapeptides not found in any of the protein sequences (Uniprot–Swiss-Prot). **(c)** Number occurrences of amino acids (regardless of positions X_1 to X_5) in the absent pentapeptides, Y, is mathematically related to their respective percentage occurrence*, P, in all known sequences by the equation $Y = 19240\,p^{-1.2706}$. **(d)** Number of occurrences of amino acids (regardless of positions X_1 to X_5) in the absent pentapeptides is independent of their respective pI and pKa values. **(e)** Number of occurrences of amino acids (regardless of positions X_1 to X_5) in the absent pentapeptides is independent of their respective molecular weights and hydropathy indices. Note that the number of occurrences of amino acids in the absent pentapeptides are shown as mean ± standard deviation ($n = 5$, for positions X_1 to X_5) in (b), (c) and (d). Also note that from (a) the number of occurrences for any given amino acid is similar, regardless of the position, and hence standard deviation of the mean is very small. * The percentage occurrence of amino acids in all known curated sequences was obtained directly from "Amino acid distribution statistics" in the "Statistics" link – "UniProtKB/Swiss-Prot UniProt release 2019_07 Jul-31, 2019".

occurrence of amino acids in the absent pentapeptides and the percentage occurrence of the amino acids in the naturally occurring primary sequences. Such a strong relationship clearly shows that the absent primary sequences (i.e. the pentapeptides in this case) and naturally present primary sequences are constrained by the same amino acid frequencies. Figures 1d, e show that the absent pentapeptide compositions are independent of other physico-chemical properties of the constituting amino acids. Therefore, it is clear that relative availability of amino acids (i.e. their stoichiometric proportions) is solely responsible for dictating which primary sequences exist and which do not. Hence, not only is the natural primary sequence space limited, but it is limited by the stoichiometric constraints on the constituting amino acids.

Implications in evolutionary and synthetic biology

Substantial work related to protein folding and function is dedicated to understanding constraints in the structural space for proteins. Here we provide the first direct insight into the constraints on natural protein sequence space. The absence of a single tetrapeptide "CQWW" (out of 160000 possible tetrapeptides) in all known protein sequence is interesting by itself. However, it extends to the fact that any sequence of the form ….CQWW…. does not exist or has not yet been found in naturally occurring (curated) sequences. Similarly, the absence of >2.6% of possible pentapeptides (83605 out of 3200000) further widens the constraints on natural sequence space. Remarkably, the constraints on the sequence space result from constraints on stoichiometric proportions of amino acids, most likely resulting from the "margin of life" discovered earlier (Mezei, 2011; Mittal et al., 2010; Mittal & Jayaram, 2011a; 2011b).

Discovery of absence small peptides (and larger sequences including them) opens up fascinating avenues in evolutionary biology. Were these peptides ever present? Is the absence of these peptides reflected at the genome level or at the translational level or only at the amino acid level? For example, CQWW can have 4 DNA sequences with 12 bases each depending on the number of codons for each residue (2 for C, 2 for Q, 1 for W; possible DNA sequences

corresponding to CQWW = 2x2x1x1 = 4). Thus, either the 12-mer DNA sequences do not exist for some reasons including but not limited to codon usage (Mittal & Jayaram, 2012). Or, if they do exist, then these are never translated into protein sequences for some reason(s) including, but not limited to, inadequate availability of the individual amino acids and/or their respective tRNAs.

In relation to the above, it is pertinent to mention that there have been a few of earlier attempts at discovering absent peptides (Navon et al., 2016; Otaki, Ienaka, Gotoh, & Yamamoto, 2005; Poznański et al., 2018; Tuller, Chor, & Nelson, 2007). However, those attempts neither verified and/or considered curation of protein sequence data (i.e. non-verified protein sequences and/or protein sequences predicted from non-coding regions were also included) nor were the absent peptides compositionally compared to sequences specifically found in nature as shown in Figure 1c here. Further, for some reasons (including, but not limited to, possible computational flaws or inclusion of non-curated protein sequence data) the absent tetrapeptide "CQWW" remain undiscovered in spite of the fact that the datasets used earlier were much smaller subsets of the protein sequence data used here (i.e. size of the datasets chosen was smaller than that in this study). Finally, in spite of finding absence of some peptides or unexpected/anomalous occurrence patterns of peptides in their respective datasets, the earlier studies completely missed investigations on compositional relationships (or lack of) between the peptides and the actual sequences in their respective datasets. Thus, the discovery of stoichiometric constraints on amino acids in protein sequences resulting in absent peptides eluded the earlier attempts.

Results reported here also open fascinating avenues in synthetic biology. Either by specific expression of the discovered peptides, or by exogenous chemical synthesis of the discovered peptides, it would be interesting to note how living systems interact with such peptides that do not apparently exist in nature. On one hand, this opens wonderful avenues to explore useful applications of the discovered peptides – one such example being possible modulation of immune responses due to varying immunogenicity of such peptides (Patel et al., 2012). On the other hand, it may be a natural warning regarding possible (toxic) negative effects of these peptides on living systems – one such example being the reported toxicity of rare/absent peptides on cell lines (Alileche & Hampikian, 2017). Regardless, the discovered library of absent peptides provides a nice complementary data set in systematic efforts towards explorations of *de novo* protein design.

Acknowledgements

AMC is grateful to IIT Delhi for fellowship support. The authors also thank IIT Delhi for providing access to the HPC facility. AM is grateful to Kusuma Trust (UK) for their generous funding support towards assisting him in establishing the teaching and research programs of the School of Biological Sciences (subsequently renamed as the Kusuma School of Biological Sciences) at IIT Delhi.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alileche, A., & Hampikian, G. (2017). The effect of Nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. *BMC Cancer*, *17*(1), 533. doi:10.1186/s12885-017-3514-z

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230. doi:10.1126/science.181.4096.223

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. Jr. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(9), 1309–1314.

Baldwin, R. L. (1995). The nature of protein folding pathways: The classical versus the new view. *Journal of Biomolecular NMR*, *5*(2), 103–109. doi:10.1007/bf00208801

Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., … Baker, D. (2016). Accurate de novo design of hyperstable constrained peptides. *Nature*, *538*(7625), 329–335.

Boyken, S. E., Benhaim, M. A., Busch, F., Jia, M., Bick, M. J., Choi, H., … Baker, D. (2019). De novo design of tunable, pH-driven conformational changes. *Science*, *364*(6441), 658–664.

Chevalier, A., Silva, D.-A., Rocklin, G. J., Hicks, D. R., Vergara, R., Murapa, P., … Baker, D. (2017). Massively parallel de novo protein design for targeted therapeutics. *Nature*, *550*(7674), 74–79.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, *357*(6379), 543–544. doi:10.1038/357543a0

Dill, K. A., & Chan, H. S. (1997). From levinthal to pathways to funnels. *Nature Structural Biology*, *4*(1), 10–19.

Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., … Baker, D. (2018). De novo design of a fluorescence-activating β-barrel. *Nature*, *561*(7724), 485–491. doi:10.1038/s41586-018-0509-0

Huang, P. S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, *537*(7620), 320–327. doi:10.1038/nature19946

Karplus, M., & Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(19), 6679–6685. doi:10.1073/pnas.0408930102

Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.-A., Bick, M. J., … Baker, D. (2019). De novo protein design by citizen scientists. *Nature*, *570*(7761), 390–394. doi:10.1038/s41586-019-1274-4

Langan, R. A., Boyken, S. E., Ng, A. H., Samson, J. A., Dods, G., Westbrook, A. M., … Baker, D. (2019). De novo design of bioactive protein switches. *Nature*, *572*(7768), 205–210. doi:10.1038/s41586-019-1432-8

Levinthal, C. (1968). Are there pathways for protein folding?. *Journal de Chimie Physique*, *65*(1), 44–45. doi:10.1051/jcp/1968650044

Levinthal, C. (1969). How to fold graciously. In *Mossbauer spectroscopy in biological systems: Proceeding of a meeting held at Allerton House*, edited by J.T.P. De Brunner and E. Munck, (pp. 22–24). Monticello, Illinois: University of Illinois Press.

Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., … Baker, D. (2018). De novo design of a non-local β-sheet protein with high stability and accuracy. *Nature Structural & Molecular Biology*, *25*(11), 1028–1034. doi:10.1038/s41594-018-0141-6

Mezei, M. (2011). Discriminatory power of stoichiometry-driven protein folding?. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 625–626. doi:10.1080/073911011010524966

Mittal, A., & Jayaram, B. (2011a). Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 443–454. doi:10.1080/073911011010524954

Mittal, A., & Jayaram, B. (2011b). The newest view on protein folding: stoichiometric and spatial unity in structural and functional diversity. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 669–674. doi:10.1080/073911011010524984

Mittal, A., & Jayaram, B. (2012). A possible molecular metric for biological evolvability. *Journal of Biosciences*, *37*(3), 573–577. doi:10.1007/s12038-012-9210-x

Mittal, A., Jayaram, B., Shenoy, S. R., & Bawa, T. S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding?. *Journal of Biomolecular Structure and Dynamics*, *28*(2), 133–142. doi:10.1080/07391102.2010.10507349

Navon, S. P., Kornberg, G., Chen, J., Schwartzman, T., Tsai, A., Puglisi, E. V., … Adira, N. (2016). Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences. *Proceedings of the National Academy of Sciences*, *113*(26), 7166–7170. doi:10.1073/pnas.1606518113

Ng, A. H., Nguyen, T. H., Gómez-Schiavon, M., Dods, G., Langan, R. A., Boyken, S. E., … El-Samad, H. (2019). Modular and tunable biological feedback control using a de novo protein switch. *Nature*, *572*(7768), 265–269. doi:10.1038/s41586-019-1425-7

Otaki, J. M., Ienaka, S., Gotoh, T., & Yamamoto, H. (2005). Availability of short amino acid sequences in proteins. *Protein Science: A Publication of the Protein Society*, *14*(3), 617–625. doi:10.1110/ps.041092605

Patel, A., Dong, J. C., Trost, B., Richardson, J. S., Tohme, S., Babiuk, S., … Kobinger, G. P. (2012). Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. *PLoS One*, *7*(8), e43802. doi:10.1371/journal.pone.0043802

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins: Two hydrogenbonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, *37*(4), 205–211. doi:10.1073/pnas.37.4.205

Poznański, J., Topiński, J., Muszewska, A., Dębski, K., J., Hoffman-Sommer, M., Pawłowski, K., & Grynberg, M. (2018). Global pentapeptide statistics are far away from expected distributions. *Scientific Reports*, *8*(1), 15178. doi:10.1038/s41598-018-33433-8

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. doi:10.1016/s0022-2836(63)80023-6

Rose, G. D., Fleming, P. J., Banavar, J. R., & Maritan, A. (2006). A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(45), 16623–16633. doi:10.1073/pnas.0606843103

Shen, H., Fallas, J. A., Lynch, E., Sheffler, W., Parry, B., Jannetty, N., … Baker, D. (2018). De novo design of self-assembling helical protein filaments. *Science*, *362*(6415), 705–709. doi:10.1126/science.aau3775

Tuller, T., Chor, B., & Nelson, N. (2007). Forbidden penta-peptides. *Protein Science*, *16*(10), 2251–2259. doi:10.1110/ps.073067607

UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res*, *47*(D1), D506–515.

Wolynes, P. G., Onuchic, J. N., & Thirumalai, D. (1995). Navigating the folding routes. *Science (New York, N.Y.)*, *267*(5204), 1619–1620. doi:10.1126/science.7886447