Received: 20 April 2019 | Revised: 9 August 2019 | Accepted: 30 August 2019
DOI: 10.1002/prot.25811

PROTEINS WILEY

RESEARCH ARTICLE

On predicting foldability of a protein from its sequence

Mihaly Mezei

Department of Pharmacological Sciences,
Icahn School of Medicine at Mount Sinai,
New York, New York

Correspondence
Mihaly Mezei, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029.
Email: mihaly.mezei@mssm.edu

Funding information
Icahn School of Medicine at Mount Sinai

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1002/prot.25811.

Abstract

Several properties of amino acid sequences corresponding to proteins that are known to fold are compared to those of randomly generated sequences and to sequences of intrinsically disordered proteins in order to find properties that distinguish folding sequences from the rest. The properties studied included helix and sheet propensities from secondary structure prediction, adjacency correlations, directionality correlations, as well as propensities of all possible triplets and quadruplets. Small differences between known folded and random sequences were observed for the adjacency and directional correlations, and significant differences were seen on the triplet and especially on the quadruplet propensities. Based on the differences in the adjacency, triplet or quadruplet propensities folding scores were defined and used to test the accuracy of foldability prediction based on these statistics. The best predictions were obtained from the quadruplet propensities.

KEYWORDS
protein foldability, residue correlation, residue propensity, secondary structure prediction

1 | INTRODUCTION AND BACKGROUND

It has been demonstrated that a randomly selected amino acid (AA) sequence is unlikely to fold: analyzing the case of 100-residue sequences, Dokholyan showed[1] that of the 20^{100} possible sequences, only about 10^{47} sequences are capable of forming a stable compact structure. That means that there are significant constraints on sequences that are capable of folding. The question thus arises: what are the features of a sequence that lend it the capability to fold?

The question of the randomness (or lack thereof) of the sequences in known proteins has been raised before.[2] Earlier studies reached conflicting conclusions about the degree of randomness of extant protein structures as reviewed by De Lucrezia et al.[3] Some of those earlier approaches looked at properties similar to those in the present work as indicated in the Methods section.

Folding into a well-defined conformation is a stricter requirement than forming a molten globule with some secondary structure (SS) elements formed. In fact, experimental studies of random sequences by LaBean et al.[4] showed that randomly generated sequences generally show evidence of forming SS elements and folding into a molten globule. Bungard et al.[5] found similar evidence of structural organization in a de novo evolved protein.

Besides the conventional approach searching for particular correlations in the data, as done in this paper as well, recent trends moved in the direction of artificial intelligence (AI or "deep learning"), like the paper by De Lucrezia et al. referred to above. The Baker Laboratory combined the Rosetta structure prediction[6] with metagenomics data integrated with AI.[7] While AI approaches can indeed work well for classifying a given data set, they are intellectually less satisfying as they do not add to the underlying science.

There have also been discussions on the existence of a folding code—a concept than can have several interpretations but it goes beyond the concept of nonrandomness. For example, Uversky[8] suggested that the right ratio of hydrophobicity to charge can be considered a code for sequences that adopt partially formed conformations. Based on tetramers, Rackovsky[9] derived a code for the SS those tetramers are likely to form. On the other hand, Ben-Naim argues that there is no such thing as a folding code, in the sense that one cannot expect to predict the structure by reading the sequence of a protein.[10]

This paper describes some attempts to find sequence clues to foldability. It is based on comparing various statistical properties of

© 2019 Wiley Periodicals, Inc.

the sequences of a large number of proteins found in the Protein Data Bank (PDB) with sequences generated randomly. The properties that were found to be able to distinguish between folding and random sequences were then used to define a foldability prediction algorithm, which was found to be largely successful in distinguishing folding and random sequences.

2 | MATERIALS AND METHODS

2.1 | Data sources

The sequence and SS of the structures in the PDB[11,12] were downloaded as the file ss.txt in 2018 from the PDB website, https://www.rcsb.org/. The 394 869 protein chains have been filtered by sequence identity and length. Keeping the larger of two chains when they have more than 90%, 70%, and 50% sequence identity, respectively, and having at least 20 residues, resulted in 47 405, 41 042, and 35 667 chains, respectively. The analyses will be run on the set filtered to 50% identity.

The filtering involved the following steps:

1. Sort the sequences by decreasing length.
2. Accept the first sequence.
3. For each subsequent sequence S_i, calculate the sequence identity with the set of accepted sequences in decreasing order until a sequence is found with greater sequence identity than the threshold. The alignment used the substitution matrix of Henikoff and Henikoff[13]; the penalty of opening a gap of −12 was −12 and of extending a gap was −1.
4. If a similar sequence is found, delete the sequence S_i. The sorting will ensure that always the shorter sequence will be dropped from a match.

In addition, before each analysis, putative HIS tags were removed. Histidine sequences of at least six residues long that are separated from either end of the sequence by less than seven residues were considered HIS tags.

For comparison, sequences were randomly generated; for some tests, intrinsically disordered protein (IDP) sequences were also used. The randomly generated sequences were either sampled from the uniform distribution (ie, each AA was selected with probability 0.05) or from the distribution of their propensity to occur in naturally existing proteins (not just the proteins in the filtered PDB set). The AA propensities were taken from Ref.14, averaged over the various organism types.

Besides the propensities averaged over organism types (referred to as generic), propensities calculated from the input set were also used. The input-based propensities were either based on the overall numbers of occurrence of each residue (referred to as data based) or calculated for each protein separately and averaged over the proteins (referred to as local). The data-based propensities calculated from the filtered sequence set were close to the generic propensities—the differences were significantly less than the differences among the propensities of the various organism types.

A total of 752 sequences describing intrinsically disordered regions (IDP set) were downloaded from the DisProt database.[15] Melting temperatures of 96 proteins were obtained from the data set of Pucci et al.[16]

2.2 | Use of secondary structure prediction algorithms

The idea here was that randomly generated sequences should result in fewer residues predicted to form SS elements than folding sequences and/or the lengths of such elements are different in the folding and nonfolding proteins. There are several algorithms that give predictions of SS from the protein sequence, starting with the algorithm of Chou–Fassman (CF)[17] and Garnier–Osguthrope–Robson (GOR),[18] followed by several more sophisticated and/or specialized ones. As an aside, it turns out that it is very difficult to reproduce these algorithms as the developers keep improving the parameters but often neglect to fully document the changes.

The SS predictions were run on the filtered sequence set from the PDB, on the IDP set, and on two randomly generated sequence sets: sampled from the uniform distribution and from the distribution defined by the AA propensities.

It is also important to emphasize that the SS predictions are used here only as a tool to look for well-defined properties that are able to discriminate between folding and nonfolding sequences. Therefore, the accuracy of the prediction method is of secondary interest for the purpose of this study.

2.3 | Use of residue adjacency statistics

It is reasonable to assume that folding biases the selection of residues that are adjacent in the sequence. To establish the magnitude of this effect for each pair (AA_1,AA_2), the ratio of the number of AA_1-$(X)_n$-AA_2 ($n \geq 0$) to the number expected if the residues are selected randomly form the AA propensity distribution of the overall sequence data set was calculated. A ratio of 1.0 indicates no correlation. A similar approach was used by Santoni et al.[19]: they looked at the statistics of the sequence distance of different AA pairs and compared it to the expected distances from random sequences generated with the natural AA propensities. They included the directionality of the protein chain in the statistics.

2.4 | Use of directionality statistics

A protein chain has a well-defined directionality. It is structurally obvious in helix structures, but it may have relevance in other parts of the protein as well. While in the present adjacency statistics, the directionality was ignored, a separate measure was used to detect possible direction effects: for each pair (AA_1,AA_2), the ratio of the number of AA_1-$(X)_n$-AA_2 sequences to the number of AA_2-$(X)_n$-AA_1 sequences was calculated ($n \geq 0$) on the filtered PDB set. A ratio of 1.0 indicates no directional effect. Separating the directionality test from the adjacency test has the advantage that the separate directionality test does not require reference propensities.

2.5 | Use of triplet and quadruplet statistics

There are 8000 different triplets and 160 000 different quadruplets that the 20 AAs can form. This means that filtered PDB set can provide good statistics for the triplets but somewhat lower quality for quadruplets.

In the first step, for all 8000 AA triplets and for all 160 000 AA quadruplets, the probability of their occurrences was determined (to be precise, approximated from their relative frequencies in the input sequence set) using the filtered PDB sequence set. Next, they were normalized by the probability of their occurrence in randomly generated sequences that were obtained from the distribution of the generic AA propensities, giving a measure for each triplet and quadruplet for their likeliness to occur in folded sequences.

In an earlier work,[20] quadruplet and quintuplet propensities were compared with the propensities expected from random sequences. For the quintuplets, a reduced AA representation was used to be able to obtain adequate precision.

2.6 | Quantifying the foldability propensity based on adjacency, triplet, and quadruplet distributions

For each sequence and construct p (pair, triplet, and quadruplet), a score SC_p was assigned as

$$SC_p = \left[\sum_{i=1}^{N} \ln\left(PN_i/PR_i\right)\right]/N$$

where N is the number of constructs in the sequence, PN_i is the probability of finding the construct i in the PDB set, and PR_i is the probability of finding the same construct in the (propensity-weighted) randomly generated set. It is expected that SC_p will be different for folding and nonfolding sequences.

For a given pair, triplet, or quadruplet, the ratio PN_i/PR_i is one if its frequency in the PDB set is what one would expect from the protein AA propensities, less than one if it occurs less frequently and more than one if it occurs more frequently. Due to taking the logarithm, zero takes the role of the separator between likely folding and likely nonfolding sequences.

The values of ln (PN_i/PR_i) were calculated for all 400 pairs, 8000 triplets, and 160 000 quadruplets once and saved. For a given sequence, the calculation is just the summation of the values corresponding to the sequence in question.

Using the precalculated ln (PN_i/PR_i) values, the foldability scores SC_r (r referring to pairs, triplets, or quadruplets) were calculated for the PDB set, the IDB set, as well as the two kinds of randomly generated sets. For each data set and each type of score, the score distributions were calculated, normalized in such a way that the area under the curve is one. Overlap between two distributions is calculated as the area of the region that is simultaneously under both curves. Clearly, it has to be between zero and one.

Denoting the score distribution over the PDB set $P_{\mathrm{PDB}}(SC)$ and over the propensity-weighted random set $P_{\mathrm{R}}(SC)$, where SC can be any of the constructs (pair, triplet, and quadruplet) or a combination thereof, the following rules were used to estimate foldability:

1. Nonfolding for sure if $P_{\mathrm{R}}(SC) > 0$ and $P_{\mathrm{PDB}}(SC) = 0$ or $SC < SC_{\mathrm{R,min}}$.
2. Folding for sure if $P_{\mathrm{PDB}}(SC) > 0$ and $P_{\mathrm{R}}(SC) = 0$ or $SC > SC_{\mathrm{PDF,max}}$.
3. Likely not to fold if $P_{\mathrm{R}}(SC)/P_{\mathrm{PDB}}(SC) > 1.0$.
4. Likely to fold $P_{\mathrm{PDB}}(SC)/P_{\mathrm{R}}(SC) > 1.0$.

However, if $0.5 < P_{\mathrm{PDB}}(SC)/P_{\mathrm{R}}(SC) < 2.0$, then the estimate is also marked as "weak."

2.7 | Accuracy of foldability predictions based on the adjacency, triplet, and quadruplet scores

To test the accuracy of predictions based on the propensity scores SC 9593, structures deposited in the PDB after the ss.txt file were downloaded and the chain information (sequence and SS) in the format of the ss.txt file have been extracted. Filtering to 50% sequence identity resulted in 4736 chains. For the comparison, 100 000 random sequences were generated with different random number seed (3579) that were used to generate the original statistics (1357).

2.8 | Software and data generated

The calculations described (filtering, conversion from .cif format, all analyses) have been performed with the Fortran program Fold, available on the website http://inka.mssm.edu/~mezei/fold. Besides the program, the website contains (a) the filtered sets of sequences used to generate the distributions/scores, ss_nr50.txt; (b) the filtered set of sequences used for the foldability predictions, ss_new_nr50.txt; and (c) a script to run all the data generation and analyses, runall.bat. Thus, data not shown in the paper can be generated—running all the analyses described in the paper takes a few minutes on a single CPU.

3 | RESULTS AND DISCUSSION

Comparison of results from the input sets filtered to different levels of similarity showed little or no appreciable differences. Therefore, results will be given only for one set, filtered at 50% sequence identity.

3.1 | Secondary structure predictions

Figure 1 shows the distribution of the predicted percentages of SS elements (helix or sheet) in the PDB sequence set, in the IDP set, and in two randomly generated (100 000 sequences of 200 residues each) sets using the uniform distribution and the AA propensity distribution, respectively, as well as for the experimental SS annotation. The predictions used the original GOR parametrization. The corresponding averages, SDs, and ranges are given in Table 1.

The distribution corresponding to the IDP set is similar to that of the PDB set albeit more noisy. There is, however, a clear difference between the distribution of percentages from PDB sequences and the sequences with residues sampled from uniform distribution but the difference mostly disappears when the random sequences are sampled from a distribution that selected each residue with probability proportional to the propensity of that AA to occur in the protein space. In hindsight, this is not surprising, considering the experimental evidence of SS formation in randomly generated sequences.[4,5]

The predicted length of the SS elements, however, was found to be different in the PDB and in both randomly generated sets. Table 2 shows the predicted average helix and sheet lengths, as well as the average predicted helix and sheet lengths in the PDB set. The closer the set is to foldability, the longer the predicted average SS element length is (although still somewhat below the experimental averages). Again, the IDP set predictions are close to that of the PDB set.

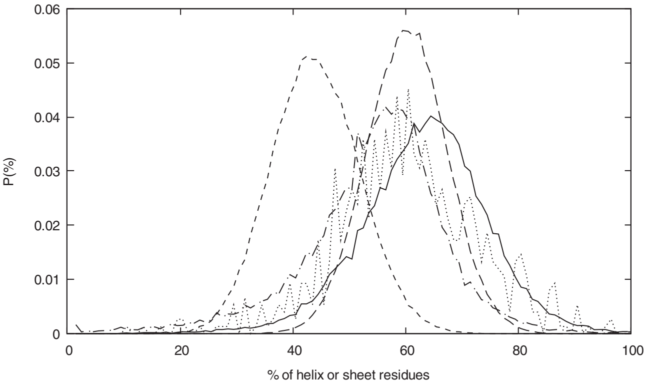

FIGURE 1 Distribution of the predicted % of residues forming helix or sheet in the PDB sequences (full line), in the IDP set (dots), in the random set following the AA propensity distribution (long dash), and in the uniformly distributed random set (short dash). The experimentally determined propensity distribution for secondary structure is also plotted (dot-dash)

Calculations were also done using a different parametrization of the GOR method as well as using versions of the CF method but they resulted in even less differences between folding and nonfolding sequences.

3.2 | Residue-residue correlation statistics

The initial calculation of the probabilities of finding two AAs in each other's vicinity presented an intriguing problem: as the distance between the correlated residues grew, the calculated correlation measure for most residue pairs converged to a value different from one. This behavior persisted no matter which propensity scheme (generic, data-based, or local) was used to normalize but were absent when (as a test for the correctness of the program) random sequences were generated using the generic propensities. This might also explain why one of the earlier works has not found any correlation.

The unreliability of the normalization scheme led to a different way of detecting adjacency correlations: instead of normalizing by putative random probabilities, the adjacency frequencies were normalized by the adjacency frequency of pairs separated by 10 residues. This factored out the reference probabilities whose use in this context was found to be unreliable. Note that the only error this normalization could introduce would be an underestimation of the magnitude of correlations.

Table 3 shows the propensities of AA pairs to be adjacent (ie, n = 0), normalized by their propensity to be 10 residues apart. Clearly,

TABLE 1 Helix and sheet propensity statistics

Source	Method	Average % of helix and sheet	SD	Minimum	Maximum
PDB	Experiment	53.4	12.5	0	99
PDB	GOR1	62.2	11.2	0	100
IDP	GOR1	59.0	12.8	10	96
wran	GOR1	59.0	7.1	29	89
uran	GOR1	42.8	7.7	14	77

Abbreviations: PDB, Protein Data Bank; IDP, intrinsically disordered protein; GOR, Garnier–Osguthrope–Robson.

TABLE 2 Helix and sheet length statistics

Source	SS element	Method	Average	SD	Minimum	Maximum
PDB	H	Experiment	9.9	6.8	1	100
PDB	S	Experiment	5.3	2.7	1	45
PDB	H	GOR1	8.3	8.0	1	100
PDB	S	GOR1	4.1	2.6	1	39
IDP	H	GOR1	8.8	9.3	1	100
IDP	S	GOR1	3.9	2.5	1	23
wran	H	GOR1	7.0	6.2	1	78
wran	S	GOR1	4.2	2.7	1	40
uran	H	GOR1	5.9	5.3	1	65
uran	S	GOR1	3.3	2.2	1	28

Abbreviations: PDB, Protein Data Bank; IDP, intrinsically disordered protein; GOR, Garnier–Osguthrope–Robson.

TABLE 3 Residue-residue adjacency propensities of the first neighbors normalized by the adjacency propensities of 10th neighbors

GLY	1.0																			
ALA	0.9	1.1																		
VAL	0.9	1.0	1.0																	
LEU	0.9	1.0	1.0	0.9																
ILE	1.0	1.0	1.0	0.9	0.9															
SER	**1.2**	1.0	1.0	1.0	1.0	1.1														
THR	1.1	1.0	1.1	1.1	1.0	1.0	1.0													
ASP	1.0	1.0	1.1	1.0	1.1	0.9	0.9	0.9												
GLU	0.9	1.0	1.0	1.0	1.0	**0.9**	0.9	1.0	1.1											
ASN	1.0	1.0	1.0	1.0	1.1	1.0	0.9	0.9	1.0	1.0										
GLN	1.0	1.1	1.0	1.1	1.0	0.9	0.9	**0.9**	1.0	1.0	1.1									
LYS	1.0	1.1	1.0	1.1	1.0	0.9	0.9	1.0	1.1	1.0	1.0	1.0								
HIS	1.0	0.9	0.9	1.0	1.1	1.1	1.0	0.9	**0.9**	0.9	1.0	**0.9**	**1.7**							
ARG	1.0	1.0	1.0	1.0	1.1	0.9	0.9	0.9	1.0	1.0	1.1	1.0	1.0	1.0						
PHE	1.0	1.0	1.0	0.9	0.9	1.1	1.1	1.1	1.0	1.0	1.0	0.9	1.1	1.0	0.9					
TYR	1.0	0.9	1.0	1.0	0.9	1.0	1.0	1.1	0.9	1.0	1.0	1.0	1.0	1.1	1.1	1.1				
TRP	0.9	0.9	0.9	1.0	0.9	1.1	1.0	1.1	1.0	1.1	**1.2**	1.0	1.1	1.0	0.9	1.0	0.9			
CYS	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0	0.9	1.0	1.0	0.9	**1.2**	1.1	1.0	1.1	0.9	**0.7**		
MET	1.0	**1.2**	1.0	0.9	0.9	1.1	1.0	1.0	1.0	1.0	1.0	1.1	**1.2**	1.0	**0.8**	**0.9**	1.0	0.9	1.0	
PRO	0.9	0.9	1.1	1.0	1.0	1.1	1.1	1.1	1.0	**1.2**	1.0	1.0	1.1	0.9	1.0	1.0	**0.8**	1.0	0.9	1.0
	GLY	ALA	VAL	LEU	ILE	SER	THR	ASP	GLU	ASN	GLN	LYS	HIS	ARG	PHE	TYR	TRP	CYS	MET	PRO

Note: Propensities >1.15 or < 1/1.15 are printed in bold face.

TABLE 4 Sequence-distance dependence of residue-residue adjacency propensities

	HIS HIS	CYS CYS	SER GLY	MET PHE	MET ALA	MET HIS	CYS HIS	PRO TRP	PRO ASN	GLN ASP	TRP GLN	HIS LYS	HIS GLU	GLU SER	MET TYR
1	1.71	0.74	1.23	0.83	1.19	1.19	1.18	0.85	1.17	0.85	1.16	0.86	0.87	0.87	0.87
2	1.64	0.81	1.07	0.88	1.03	1.05	1.01	0.98	1.05	0.87	1.13	0.94	0.90	0.93	0.95
3	1.46	0.56	1.01	0.98	1.02	1.05	1.03	0.96	1.06	0.98	1.05	0.93	0.94	1.00	0.96
4	1.39	0.99	1.03	0.97	1.02	1.01	1.07	0.98	1.03	0.92	1.11	0.95	0.97	0.96	0.95
5	1.23	0.96	1.00	0.95	1.02	1.03	1.01	0.95	1.05	0.95	1.07	0.99	0.98	0.96	0.96
6	1.10	0.99	1.00	0.92	1.00	1.02	1.02	0.94	1.04	0.95	1.04	0.98	0.96	0.96	0.99
7	1.08	0.96	1.01	1.00	1.01	1.02	1.01	1.00	1.03	0.98	1.04	0.98	0.97	0.99	0.98
8	1.03	0.99	1.00	0.99	1.02	1.03	1.02	0.99	1.01	1.00	1.03	0.97	0.99	0.99	0.99
9	1.01	0.99	1.00	0.95	1.02	1.05	1.00	0.99	1.00	0.99	1.03	1.00	0.97	0.99	0.98
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.0

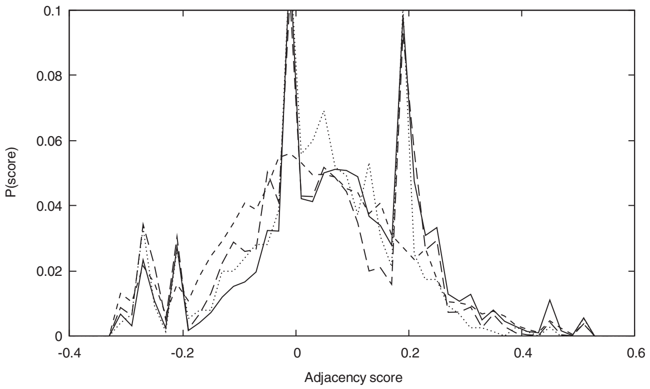

FIGURE 2 Adjacency score distribution for the nonredundant PDB sequences (full line), for the propensity-based random sequences (long dashes), for the uniformly distributed random sequences (short dashes), and for the IDP sequences (dots)

for most AA pairs, there is no noticeable correlation. Comparison of the same table calculated on the first and last half of the data set indicates that the precision of the values in the table is of the order of 10%. Accordingly, adjacency propensities exceeding 1.15 or below 1/1.15 are shown in bold face.

The decay of the correlations seen with sequence distance is another indication of the significance of the data. For the AA pairs, the sequence-distance dependence of neighborhood propensities (ie, $n > 0$) where the adjacency propensity exceeded 1.15 or was below 1/1.15 is shown in Table 4 as a function of n. While the difference in the presentation of the results does not allow detailed comparison with the results of Santoni et al.,[19] in both studies, histidine was shown to be involved in the most and largest correlations.

The distribution of scores SC_p (defined above in Section 2.6) for the PDB and IDP sets and for the randomly generated sets (again, 100 000 sequences of 200 AA each) are shown in Figure 2. The distributions are rather noisy, and there is no clear separation between any of these distributions although at the high-score range, the PDB distribution tops all the others.

3.3 | Directionality statistics

The ratios of forward and backward neighborhood propensities are shown in Table 5. Ratios exceeding 1.15 or below 1/1.15 are shown with bold face. For most residue pairs, no noticeable asymmetry is found. It is probably not surprising that among the residue pairs with significant asymmetry, PRO is the most prominent since proline is the only AA that has no side-chain torsional freedom. The largest asymmetry (≥ 1.15 or ≤ 1.15, shown in bold face in Table 5) was found for the residue pairs MET-HIS (0.56), PRO-GLU (1.62), PRO-CYS (0.72), MET-THR (1.32), MET-TRP (0.76), PRO-HIS (0.77), PRO-ILE (0.78), ASN-GLU (0.78), PRO-TRP (1.26), PRO-ASN (0.80), PRO-GLY (1.25), MET-LYS (1.25), TRP-LEU (1.23), PRO-MET (0.82), HIS-SER (0.83), PRO-THR (0.83), HIS-LYS (0.83), LYS-GLY (0.83), MET-ASN (1.19), HIS-GLU (0.84), VAL-GLY (0.85), TYR-PHE (1.17), PHE-ARG (0.86), CYS-ILE (0.86), TRP-GLY (0.86), and CYS-TYR (0.87)—the numbers in parenthesis give the forward/backward propensity ratios.

3.4 | Triplet and quadruplet statistics

Most triplet and quadruplet propensities are significantly different from the propensities expected from the overall AA propensities. Table 6 shows the triplets for which $|\ln (PN_i/PR_i)| > 1$ and Table 7 shows the quadruplets for which $|\ln (PN_i/PR_i)| > 3.2$. Again, the frequent occurrence of histidines is noticeable (even though HIS tags were removed). Furthermore, there are 672 quadruplets that are missing from the filtered PDB data set. For those quadruplets, PN_i was set to 0.5/160 000.

The distributions of scores for the PDB set and for the randomly generated set (again, 100 000 sequences of 200 AA each) are shown in Figures 3 and 4 for the triplets and quadruplet scores, respectively. The fact that the distribution of the PDB sequence scores is shifted toward positive scores indicates that certain quadruplets are significantly more likely to occur in folded sequences than what would follow from the AA propensities. Conversely, the fact that the distribution of the randomly generated sequence scores is shifted toward negative scores indicates that such sequences contain significantly more of the triplets

TABLE 5 Adjacency asymmetry propensities

GLY	1.0																			
ALA	1.1	1.0																		
VAL	**0.8**	1.0	1.0																	
LEU	1.0	1.0	1.0	1.0																
ILE	0.9	1.0	1.0	1.0	1.0															
SER	1.0	1.0	0.9	1.0	0.9	1.0														
THR	1.0	1.0	1.0	1.0	0.9	1.0	1.0													
ASP	1.1	1.0	1.0	1.0	1.0	0.9	0.9	1.0												
GLU	1.0	1.0	1.0	1.0	1.1	0.9	1.0	0.9	1.0											
ASN	1.1	1.1	1.0	1.0	1.0	1.0	1.0	1.0	**0.8**	1.0										
GLN	1.0	1.0	1.1	1.0	1.0	0.9	1.0	1.1	0.9	1.0	1.0									
LYS	**0.8**	1.0	1.0	0.9	1.1	0.9	1.1	1.1	1.0	1.1	1.0	1.0								
HIS	1.0	0.9	1.0	1.1	1.0	**0.8**	1.0	1.0	**0.8**	1.0	1.0	**0.8**	1.0							
ARG	0.9	0.9	1.0	1.0	1.1	0.9	1.0	1.1	1.0	1.1	1.0	1.0	1.0	1.0						
PHE	1.0	0.9	1.0	1.1	1.0	1.1	1.0	1.0	1.0	1.0	1.1	1.0	0.9	**0.9**	1.0					
TYR	0.9	1.0	1.0	1.1	1.0	1.0	1.0	0.9	1.0	1.0	1.1	0.9	0.9	1.0	**1.2**	1.0				
TRP	**0.9**	0.9	1.0	**1.2**	1.1	1.0	1.0	0.9	0.9	1.0	1.1	1.1	1.0	1.0	1.0	1.0	1.0			
CYS	1.1	0.9	0.9	1.0	**0.9**	1.1	0.9	1.0	1.0	0.9	1.1	1.1	1.0	1.1	0.9	**0.9**	0.9	1.0		
MET	1.0	1.1	1.1	1.1	1.1	1.1	**1.3**	1.1	1.0	**1.2**	1.0	**1.2**	**0.6**	1.1	1.0	0.9	**0.8**	1.1	1.0	
PRO	**1.3**	1.1	1.0	0.9	**0.8**	1.0	**0.8**	1.0	**1.6**	**0.8**	1.1	0.9	**0.8**	1.0	1.1	1.0	**1.3**	**0.7**	**0.8**	1.0
	GLY	ALA	VAL	LEU	ILE	SER	THR	ASP	GLU	ASN	GLN	LYS	HIS	ARG	PHE	TYR	TRP	CYS	MET	PRO

Note: Propensities >1.15 or < 1/1.15 are printed in bold face.

TABLE 6 Absolute and relative triplet propensities with |ln (PN_i/PR_i)| > 1

Sequence	1000*PN_i	ln (PN_i/PR_i)
GLY-SER-HIS	0.1134	1.108
GLY-HIS-MET	0.0331	1.117
ALA-ALA-ALA	0.3503	1.092
SER-HIS-MET	0.0417	1.873
ASP-ASP-ASP	0.1053	1.041
GLU-GLU-GLU	0.1770	1.012
GLN-GLN-GLN	0.0450	1.034
HIS-HIS-MET	0.0114	1.247
HIS-MET-ALA	0.0365	1.157
TYR-PHE-GLN	0.0580	1.483
TRP-PRO-CYS	0.0117	−1.236
HIS-HIS-HIS	0.0106	3.302

TABLE 7 Absolute and relative quadruplet propensities with |ln (PN_i/PR_i)| > 3.2

Sequence	1000*PN_i	ln (PN_i/PR_i)
ALA-PHE-MET-CYS	0.1477	−3.386
ALA-CYS-SER-TRP	0.1419	−3.346
LEU-TYR-PHE-GLN	0.5778	3.548
ILE-TRP-MET-VAL	0.1275	−3.239
SER-ILE-CYS-TRP	0.1273	−3.237
THR-CYS-THR-MET	0.1332	−3.282
ASN-MET-MET-THR	0.1569	−3.446
HIS-HIS-HIS-TRP	0.0141	3.785
ARG-GLY-SER-HIS	0.5341	3.331
PHE-CYS-TYR-ASN	0.1580	−3.453
TYR-PHE-GLN-GLY	0.3710	3.317
TRP-SER-HIS-PRO	0.1098	3.430
GLY-SER-HIS-MET	0.2671	4.477
ILE-HIS-HIS-HIS	0.0670	4.027
HIS-HIS-HIS-MET	0.0250	4.578
ARG-GLY-CYS-PHE	0.2683	−3.983
HIS-HIS-HIS-HIS	0.0233	6.738

and quadruplets that are less likely to occur in folded sequences than what would follow from the AA propensities.

While there is a significant overlap between the distributions of the PDB set and the randomly generated sets, both the triplet and the quadruplet score distributions on the random sets are quite distinct from that of the PDB set. The difference is slightly less when the random sequences are sampled from the AA propensity distribution. The differences are larger for the quadruplet score distributions: the overlap between the triplet PDB and propensity-sampled random distribution is 0.307 while the corresponding overlap between the quadruplet distributions is only 0.191 (for identical distributions, the overlap would be 1.0).

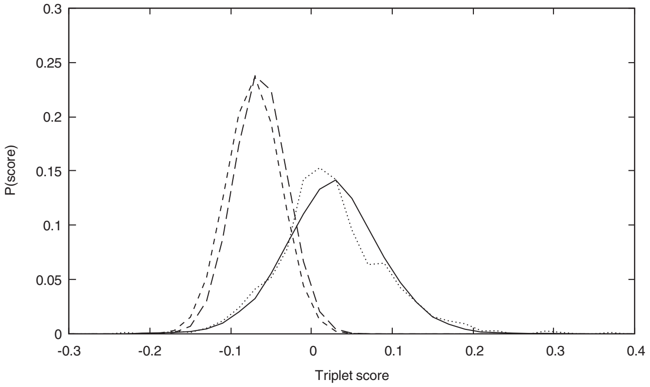

FIGURE 3 Triplet score distribution for the nonredundant PDB sequences (full line), for the propensity-based random sequences (long dashes), for the uniformly distributed random sequences (short dashes), and for the IDP sequences (dots)

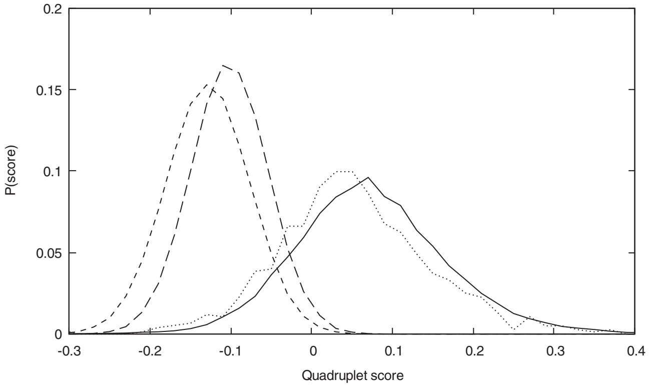

FIGURE 4 Quadruplet score distribution for the nonredundant PDB sequences (full line), for the propensity-based random sequences (long dashes), for the uniformly distributed random sequences (short dashes), and for the IDP sequences (dots)

TABLE 8 Triplet score statistics

Data set	<Triplet score>	SD	Number of sequences
PDB	0.047	0.062	35 668
IDP	0.044	0.067	752
W-random	−0.043	0.032	100 000
U-random	−0.051	0.034	100 000

Abbreviations: PDB, Protein Data Bank; IDP, intrinsically disordered protein.

The IDP triplet and quadruplet score distributions are also shown in Figures 3 and 4. They are both close to the PDB distribution, but are slightly shifted toward the random distributions. Furthermore, Tables 8 and 9 show the average triplet and quadruplet scores, respectively, for the four different sets. As expected, the IDP scores are slightly lower than the PDB scores for both the triplet and quadruplet scores, indicating weaker tendency for folding.

Table 9 also shows the correlation of the triplet and quadruplet scores. On all four data sets, it is of the order of 0.9. This suggests that

TABLE 9 Quadruplet score statistics

Data set	<Quadruplet score>	SD	Number of sequences	Triplet-quadruplet score correlation
PDB	0.094	0.097	35 668	0.94
IDP	0.070	0.100	752	0.97
W-random	−0.082	0.048	100 000	0.94
U-random	−0.109	0.052	100 000	0.89

Abbreviations: PDB, Protein Data Bank; IDP, intrinsically disordered protein.

combining the triplet and quadruplet scores would not improve the separation of distributions. In fact, combining (adding) the triple and quadruplet scores, the overlap between the PDB scores and the propensity-sampled random distribution scores is 0.225—it is larger than the overlap with the quadruplet scores only.

3.5 | Conformation dependence of the triplet and quadruplet statistics

As discussed in the introduction, the current study is aimed at characterizing sequences without the knowledge of their structure (if any). However, it is also important to note that the AA distributions are different for different secondary-structure elements. While a detailed study of such distinctions is beyond the scope of this work, a few comparisons were carried out using our triplet and quadruplet scores based on SS-specific distributions.

In the first step, triplet and quadruplet distributions were calculated separately for segments that were annotated as helices, sheets, and loops (defined here as segments not annotated as helix or sheet). Next, the distributions of the scores from the same sequence set as used in the earlier studies (Figures 3 and 4) and from the random (sampled from the propensity distribution) set were calculated. The shapes of these distributions were similar to those in Figures 3 and 4. Since it was shown that using the scores based on the triplet and quadruplet scores from the full set were able to differentiate between random and folding sequences to a good degree as demonstrated by the small overlap between the corresponding distributions, the results of the SS-dependent scores are also presented in terms of distribution overlaps.

The overlaps between the distributions based on the three SS types are shown in Table 10. It is somewhat surprising that differences among triplet distributions are rather small (ie, the overlaps are so large). For quadruplets, the overlaps are less, especially between helices and sheets.

The overlaps between the distributions based on the SS-specific PDB set and the propensity-based random set are shown in Table 11. There is no significant difference among the overlaps of three SS distribution with the random set as the full PDB set. It is thus no surprise that both the triplet and quadruplet SS distributions have similar overlaps with the random set as the full PDB set.

TABLE 10 Overlaps between the various triplet and quadruplet score distributions using SS-based statistics on the PDB sequence set

Statistics sources		Triplet overlap	Quadruplet overlap
Helix	Sheet	0.74	0.23
Helix	Loop	0.76	0.50
Sheet	Loop	0.69	0.60

TABLE 11 Overlaps between triplet and quadruplet score distributions limited to different SS elements of the PDB set and the propensity-based random set

Statistics source	Triplet overlap	Quadruplet overlap
Helix	0.26	0.22
Sheet	0.29	0.23
Loop	0.25	0.20

3.6 | Melting temperature calculations

Since scores based on the triplet or quadruplet distribution were shown to display different behavior for folding and nonfolding sequences, the question arose if these scores can also be used as an indicator of stability. To address this question, the melting temperatures of the 96 proteins[16] were correlated with their triplet scores. Both Pearson and Spearman (rank) correlations were calculated, resulting in correlation coefficients of 0.12 and 0.10, respectively. Given the high correlation between triplet and quadruplet scores, correlation with the quadruplet scores cannot be significantly higher. This result suggests that neither the triplet nor the quadruplets score is a useful measure for the characterization of the stability of folded proteins.

3.7 | Foldability predictions based on the score distributions

As the triplet and quadruplet score distributions showed significant separation between the folding and random sequences (unlike the adjacency score distribution), the results of prediction calculations are presented only using the triplet and quadruplet score distributions. Tables 12 and 13 present the foldability predictions based on the triplet and quadruplet score distributions, respectively. The predictions using the quadruplet distributions are found to be slightly more reliable than those based on the triplet distributions. Also, the predictions on the random sequence sets that were generated with the experimental AA propensities predicted slightly more sequences to be foldable than the set using the uniformly random AA distribution. This is in accordance with the conclusion of the SS prediction study's result (Section 3.1) suggesting that the experimental AA distribution is one contribution to foldability.

TABLE 12 Folding predictions using the triplet scores

	New PDB set		Uniform random		Propensity-weighted random	
Sure folded	652	13.8%	4	0.0%	0	0%
Sure random	0	0%	0	0%	0	0%
Guess folded	2934	62.0%	6346	6.3%	9412	9.4%
Guess random	1149	24.3%	93 650	93.7%	90 588	90.6%
"Weak" guess	7	0.1%	310	0.3%	124	0.1%

TABLE 13 Folding predictions using the quadruplet scores

	New PDB set		Uniform random		Propensity-weighted random	
Sure folded	1347	26.3%	8	0.0%	2	0%
Sure random	0	0%	0	0%	0	0%
Guess folded	2508	53.0%	3956	4.0%	4.255	4.3%
Guess random	980	20.7%	96 036	96.0%	95.745	95.7%
"Weak" guess	275	5.8%	2189	2.2%	5323	5.3%

Tests were also run trying (a) using the adjacency score distribution but limiting the score calculations for residue pairs that show significant separation between folding and random or (b) combination of predictors (either with arithmetic or with geometric averaging) but no improvement was obtained in either case.

4 | CONCLUSIONS

The analysis of the folded sequences in the PDB yielded several properties that are significantly different in the PDB set from randomly generated sets. The SS prediction test confirmed the importance of the sequence following the known AA propensities but was found to be unable to distinguish foldable and randomly generated but following the experimental AA distribution. Small but significant differences were seen in the residue adjacency propensities; smaller but still significant differences were found in the asymmetry of residue pair propensities. Much larger differences were obtained with the distribution of triplets and quadruplets prompting the definition of a propensity score; distribution of these scores for folding and randomly generated sequences were utilized to formulate a prediction whether a given sequence is likely to fold. The predictions were largely successful. On the other hand, an attempt to use the propensity score to correlate it with stability was unsuccessful.

Comparison of the properties of randomly generated sequences using the uniform and the AA-propensity distributions shows that sequences whose composition follows the AA-propensity distribution are more similar to the folding sequences than sequences based on the uniform AA distribution. This difference is most evidenced in the results from the SS prediction algorithm. This leads to the conclusion that the nonuniform AA distribution is one contributor to foldability. However, other results in this paper also show that simply following the AA propensity distribution is unlikely to guarantee foldability.

The present work tested differentiating between folding and randomly generated sequences using only the triplet or quadruplet distribution. There are other properties that showed differences between folding and randomly generated sequences but they were not included in the folding prediction. Future work may find a way to use these additional properties to improve the folding/nonfolding prediction.

ACKNOWLEDGMENTS

Prof. George Rose is thanked for helpful suggestions as well as for critical reading of an early version of the manuscript. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

CONFLICT OF INTEREST

The author declares no potential conflict of interest.

ORCID

Mihaly Mezei https://orcid.org/0000-0003-0294-4307

REFERENCES

1. Dokholyan N. Protein Designability and Engineering. In: PEB JG, ed. Hoboken, NJ: Wiley-Blackwell; 2009.
2. Weiss O, Herzel H. Correlations in protein sequences and property codes. *J Theor Biol*. 1997;190:341-353.
3. De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G. Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. *PLOS One*. 2012;7:e36634.
4. LaBean TH, Butt TR, Kauffman SA, Schultes EA. Protein folding absent selection. *Genes (Basel)*. 2011;2:608-626.
5. Bungard D, Copple JS, Yan J, et al. Foldability of a natural de novo evolved protein. *Cell*. 2017;24:1687-1696.
6. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. *Methods Enzymol*. 2004;383:66-93.
7. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science*. 2017;355:294-298.
8. Uversky VN. Cracking the folding code. Why do some proteins adopt partially folded conformations, whereas other don't? *FEBS Lett*. 2002; 514:181-183.

9. Rackovsky S. On the nature of protein folding code. *Proc Natl Acad Sci U S A*. 1993;90:644-648.
10. Ben-Naim A. *Myths and Verities in Protein Folding Theories*. Singapore: World Scientific; 2015.
11. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*. 2003;10(12):980.
12. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28:235-242.
13. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89:10915-10919.
14. Gaur RK. Amino acid frequency distribution among eukaryotic proteins. *IIOAB Journal*. 2014;5:6-11.
15. Piovesan D, Tabaro F, Marco IM, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res*. 2016;45: D219-D227.
16. Pucci F, Bourgeas R, Rooman M. High-quality thermodynamic data on the stability changes of proteins upon single-site mutations. *J Phys Chem Ref Data*. 2016;45:023104.
17. Chou PI, Fassman GD. Prediction of protein conformation. *Biochemistry*. 1974;13:211-222.
18. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*. 1978;120:97-120.
19. Santoni D, Felici G, Vergni D. Natural vs. random protein sequences: discovering combinatorics properties on amino acid words. *J Theor Biol*. 2016;391:13-20.
20. Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. *Bioinformatics*. 2010;26:310-318.

How to cite this article: Mezei M. On predicting foldability of a protein from its sequence. *Proteins*. 2020;88:355–365. https://doi.org/10.1002/prot.25811