Received: 15 January 2020 | Revised: 26 March 2020 | Accepted: 7 May 2020

DOI: 10.1002/prot.25900

RESEARCH ARTICLE

PROTEINS WILEY

A protein sequence fitness function for identifying natural and nonnatural proteins

Rahul Kaushik | **Kam Y. J. Zhang**

Laboratory for Structural Bioinformatics,
Center for Biosystems Dynamics Research,
Yokohama, Kanagawa, Japan

Correspondence
Kam Y. J. Zhang, Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, 1-7-22 Suehiro, Yokohama, Kanagawa 230-0045, Japan.
Email: kamzhang@riken.jp

Funding information
Japan Society for the Promotion of Science, Grant/Award Number: 18H02395

Abstract

The infinitesimally small sequence space naturally scouted in the millions of years of evolution suggests that the natural proteins are constrained by some functional prerequisites and should differ from randomly generated sequences. We have developed a protein sequence fitness scoring function that implements sequence and corresponding secondary structural information at tripeptide levels to differentiate natural and nonnatural proteins. The proposed fitness function is extensively validated on a dataset of about 210 000 natural and nonnatural protein sequences and benchmarked with existing methods for differentiating natural and nonnatural proteins. The high sensitivity, specificity, and percentage accuracy (0.81%, 0.95%, and 91% respectively) of the fitness function demonstrates its potential application for sampling the protein sequences with higher probability of mimicking natural proteins. Moreover, the four major classes of proteins (α proteins, β proteins, α/β proteins, and α + β proteins) are separately analyzed and β proteins are found to score slightly lower as compared to other classes. Further, an analysis of about 250 designed proteins (adopted from previously reported cases) helped to define the boundaries for sampling the ideal protein sequences. The protein sequence characterization aided by the proposed fitness function could facilitate the exploration of new perspectives in the design of novel functional proteins.

KEYWORDS

amino acid propensity, protein foldability, computational protein design, natural proteins, protein sequence space, scoring of protein designs

1 | INTRODUCTION

Delineating the connection between protein sequence and its structure is one of the most persuasive, debatable, and unresolved affairs in the field of computational structure biology.[1-3] In the last two decades, the concepts of delicate contribution of natural selection and the modest alteration by evolution in random copolymers in emergence of known proteins is extensively discussed and argued for its significance in origin of Life.[4-8] The specificity of existent natural proteins to encrypt unique protein structures is restricted within a limited number of folds (1457 protein folds) which poses another scientific challenge of quantifying the protein designability of sampled sequences into the existing folds. Creating a protein to perform a predefined or novel biological function(s) is often considered as protein design. In general, protein design is formulated as composing an amino acid sequence which should ideally fold into a stable structure, meant to perform some biological function(s).[9-13] The goals of protein design include either desired optimization of certain characteristics such as stability, solubility, and binding affinity or designing entirely novel sequences resulting into novel structures or attaining remedial or industrial utilities.[14-17] A primary and a very crucial step of novel protein design involves the computational identification or generation of potential protein sequences having a considerably high probability of mimicking the naturally occurring proteins and eventually folding into a compact structure.[18,19] Some of the earlier studies measured the degree of randomness in the known protein sequences to explore the logical explanations with conflicting inferences for constrained available sequence space.[5,7,8,19-22] Some of the previous computational studies of random sequence proteins argued over the extent of variability among natural protein sequences and random protein sequences.[23-25] In protein design regimes, folding into a distinct conformation is foremost requirement. It is believed that the randomly generated and *de-novo* evolved protein sequences tend to form a molten globule state with marginal secondary structural elements.[18,26,27] However, most of the recent approaches fortified with deep learning and artificial intelligence have contributed significantly in classifying the dataset[6,7,19,28,29] but without exploring the underlying science.

For the 100 residue long protein sequences, the theoretical sequences space of astronomically staggering ~10^{130} proteins (20^{100} combinations) in contrast to the infinitesimal fraction of naturally existing proteins. For instance, when a nonredundant dataset of all available protein sequences in the UniProt database (~22 million sequences, excluding predicted, and uncertain proteins) is analyzed, only ~10^9 unique stretches of 100 residues could be extracted. This huge decline in the number of available compact structures and unique 100 residues polypeptides substantiate the possibility of some underlying protein signatures at sequence level that lend the protein with potential of imitating the natural proteins and fold into stable structure. Some of the previous studies have utilized the concept of neighboring effect of amino acid residues in dictating protein secondary and tertiary structures.[30-33]

Here, we describe a sequence and secondary structure-based fitness scoring function to identify potentially foldable/designable protein sequences by differentiating them from nonnatural protein sequences. The presented fitness function implements the competency scores derived from sequences and corresponding secondary structures of well-characterized known protein domains (natural proteins, NP) and computationally generated nonnatural protein sequences with natural amino acid compositions (NNP-NC) and with uniform amino acid compositions (NNP-UC). The scoring function classifies a query protein sequence into foldable (natural protein) or non-foldable (nonnatural and/or random protein) depending on its competency scores compared with natural and nonnatural protein sequences.

2 | MATERIALS AND METHODS

For the development of the scoring function, the datasets of natural protein (NP) sequences (adopted from known protein domains) and computationally generated nonnatural protein (NNP-NC and NNP-UC) sequences are compiled.

2.1 | Dataset compilation

The protein sequences and corresponding secondary structures of all known protein domains in the latest stable release of the SCOPe database,[34] SCOPe 2.07 are extracted which comprises 274 230 protein domains. These domains are subjected to clustering at 100% sequence identity level using CD-HIT[35] to filter out the redundant proteins in the dataset. Post-clustering, resulting 77 280 domains are further screened for the presence of non-standard amino acid residues, missing residues (except for N and C terminals), domains having less than 50 residues, domains having more than 700 residues, or membrane protein domains. These filters resulted in a dataset of 58 758 globular protein domains as depicted in the Figure S1. The 100% sequence identity level filter is used to ensure the inclusion of a maximum number of possible triplets of amino acid residues and corresponding secondary structural elements. However, sequence identity filters at 80%, 60%, and 40% sequence identity levels are also used to explore the possibilities. A significant decline in the available number of combinations of triplets of amino acid residues and corresponding secondary structure is observed. The statistics related to availability of combinations of triplets is shown in Figure S2 and Supplementary Note I. The dataset corresponding to these protein sequences is referred as natural proteins (NP) dataset hereafter, as it is derived from naturally existing known proteins.

Similarly, a dataset of 65 000 proteins having sequence length varying from 50 to 700 residues is generated computationally restraining the amino acid compositions adopted from UniProtKB.[36] Since the dataset of computationally generated protein sequences is constrained to the same amino acid composition as naturally existing proteins, it is referred to as nonnatural protein dataset with natural distribution of amino acid compositions (NNP-NC). The "*makeprotseq*" module of EMBOSS[37] is used for computationally generating these protein sequences. As a cautionary measure, the NNP-NC dataset is also subjected to clustering at 100% sequence identity level to avoid the sequence redundancy. However, it is observed that these computationally generated sequences did not have any redundancy at 100% sequence identity.

2.2 | Extraction of secondary structural information

For the selected natural protein (NP) dataset, the secondary structural information at an individual residue level for each protein is extracted using the standalone version of STRIDE secondary structure assignment program.[38] The 8-class secondary structure assignment of STRIDE is converted into 3-class secondary structure assignment for further processing. In this conversion, the 3_{10} helices (G), π-helices (I), and 4-turn helices (H) are grouped as helices, the extended strands in β-sheet conformations (E) and isolated β-bridges (B) are pooled together as strands (E), and the hydrogen bonded turns (T), coils (C) and bends (S) are bundled as loops (C). For secondary structural information corresponding to nonnatural proteins with natural AA compositions (NNP-NC) dataset, secondary structure prediction using standalone version PSIPRED (PSIPRED 4.02) is performed.[39] Considering the current state of the art for protein secondary structure prediction, PSIPRED is reported to deliver a reasonably high accuracy and thus used in present study. It is worth noting that the PSIPRED failed to predict any secondary structure for 3862 proteins. These proteins are discarded from any further processing. A sub-dataset of 58 758 proteins is selected from the nonnatural proteins (NNP-NC) dataset (out of 61 138 proteins with predicted secondary structure). This led into a total of 117 516 proteins sequences, comprising 58 758 proteins each in natural proteins (NP) and nonnatural proteins with natural AA compositions (NNP-NC) datasets. To examine the differences in amino acid neighbor preferences in different secondary structures for computationally generated sequences NNP-NC, we derived the conditional probabilities of triplets using natural protein and nonnatural protein (NNP-NC) sequences and corresponding secondary structures.

2.3 | Classifying into reference and test datasets

The natural proteins (NP) and nonnatural proteins (NNP-NC) datasets are randomly separated into two parts each as reference dataset of 41 132 proteins and test dataset of 17 626 proteins (reference = 70% and test = 30% of 58 758 proteins). This resulted into four sub-datasets, viz. natural proteins reference dataset (comprising 41 132 proteins), natural proteins test dataset (comprising 17 626 proteins), nonnatural proteins (NNP-NC) reference dataset (comprising 41 132 proteins) and nonnatural proteins (NNP-NC) test dataset (comprising 17 626 proteins). The reference datasets are used for deriving a conditional probability-based statistical model, leading to competency scores of tripeptides and the test datasets are used for testing the efficiency of competency scores in distinguishing the natural protein (NP) and nonnatural protein (NNP-NC) sequences.

2.4 | Compiling sequence-based scoring libraries

For all the protein sequences in the natural proteins reference dataset, tripeptides frequencies are calculated for all possible 8000 combinations. Also, individual amino acid residues occurrence frequencies are calculated from natural proteins reference dataset. It may be noted that the natural protein reference dataset represents all the possible combinations at tripeptides level sufficiently, encompassing more than 8 million tripeptides (depicted in Figure S3). The residue occurrence frequencies and tripeptide frequencies are further used for calculating tripeptide conditional probabilities using Equation (1). Notably, the conditional probability calculated in Equation (1) considers forward (C-terminal) and backward (N-terminal residue) neighborhoods of the central residue. Also, this consideration takes care of directionality in the tripeptides as $P(Y_M|X_NZ_C)$ is not same as $P(Y_M|Z_NX_C)$. So, it may be considered that the conditional probabilities of tripeptides calculated in Equation (1) is inclusive of their residue-based adjacency and directionality statistics.

$$P(Y_M|X_NZ_C) = \frac{P(XYZ)}{P(Y)}, \tag{1}$$

where X, Y, and Z belong to any of the standard amino acid residues; $P(Y_M|X_NZ_C)$ is the conditional probability of residue "Y", given a residue "X" on its N-terminal and a residue "Z" on its C-terminal; $P(XYZ)$ is the probability of tripeptide "XYZ"; and P(Y) is the probability of residue "Y".

The conditional probabilities of all tripeptides as calculated using Equation (1) are further used to compute a percentage sequence competency score (CS-Score) at individual residue level by normalizing the conditional probabilities with the maximum conditional probability in all combinations of tripeptides. The CS-Score is calculated for the middle residue in a tripeptide considering one adjacent residue on its either side (one toward N-terminal and one toward C-terminal) using Equation (2).

$$\mathrm{CS-score}\,(\mathrm{X_N Y_M Z_C}) = 100\left(\frac{P(Y_M|X_NZ_C)}{P_{\max}(AA_M|AA_NAA_C)}\right), \tag{2}$$

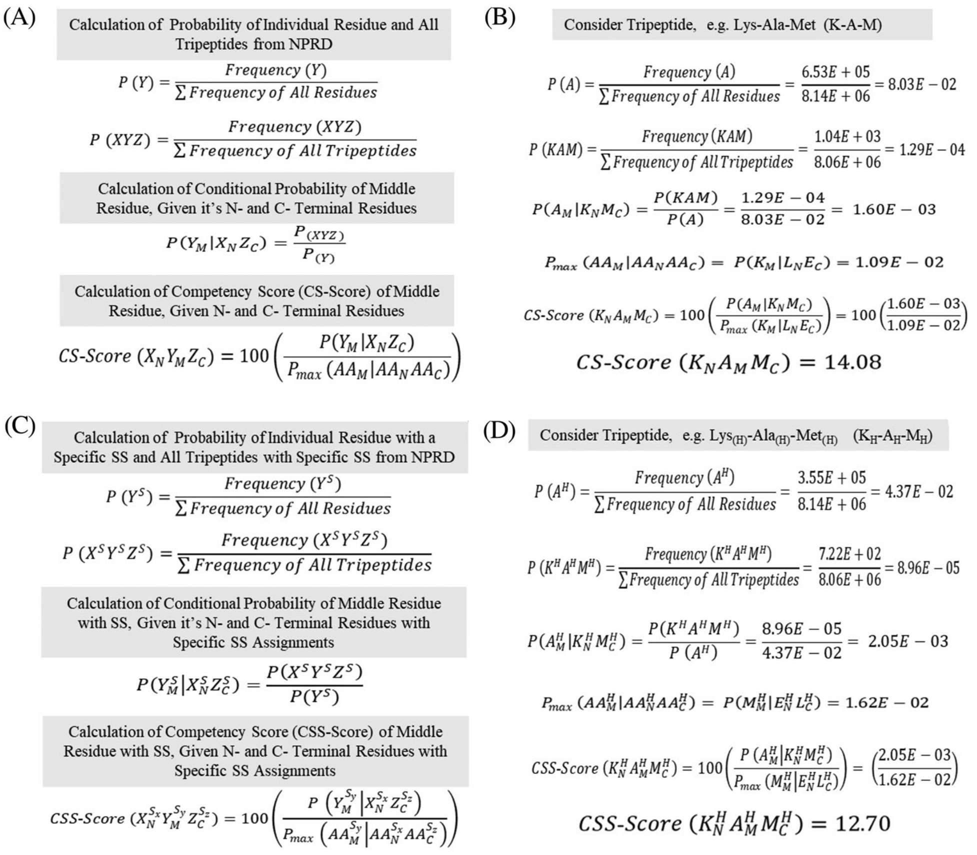
where CS-score ($\mathrm{X_NY_M\ Z_C}$) is the competency score of middle residue "Y" given residues X and Z at its N-terminal and C-terminal, respectively; $P(Y_M|X_NZ_C)$ is the conditional probability of residue "Y", given a residue "X" on its N-terminal and a residue "Z" on its C-terminal (as computed in Equation (1)); and $\mathrm{P_{max}}$ ($\mathrm{AA_M|AA_NAA_C}$) is the maximum conditional probability in all 8000 tripeptides.

The CS-Scores derived from Equation (2) resulted in 400 values for an individual residue, accounting for the occurrence of any of the 20 amino acid residues on either side. The overall flow of computation of CS-Scores is depicted in Figure 1A,B, with an example tripeptide, Lys-Ala-Met. These scores are used to evaluate the overall competence of protein sequences as discussed in results section.

2.5 | Compiling sequence and secondary structure based scoring libraries

As mentioned above, the secondary structural information at 3-class levels is compiled from natural proteins dataset. The tripeptide frequencies along with corresponding secondary structure assignments (Helix (H), Strand (E), and Coils (C)) are derived from the natural proteins reference datasets for all possible combinations, that is, 216 000 combinations ($20^3 \times 3^3$). It may be noted that all the possible combinations could not be observed in the natural protein reference dataset as seven out of 27 secondary structure combinations are practically not possible, viz. HEH, HEC, EHE, EHC, CHC, CHC, CEH. The secondary structure directed tripeptide frequencies are used to derive the probability of each available combination. Also, for all the individual residues with their secondary structure assignment (20 × 3 combinations), probabilities are calculated. The tripeptide and individual residue probabilities are further used for calculating secondary structure directed tripeptides conditional probabilities using Equation (3).

$$P\left(Y_M^{S_y}|Y_N^{S_x}Y_C^{S_z}\right) = \frac{P\left(X^{S_x}Y^{S_y}Z^{S_z}\right)}{P\left(Y^{S_y}\right)}, \tag{3}$$



(A)

Calculation of Probability of Individual Residue and All Tripeptides from NPRD

$$P\ (Y) = \frac{Frequency\ (Y)}{\sum Frequency\ of\ All\ Residues}$$

$$P\ (XYZ) = \frac{Frequency\ (XYZ)}{\sum Frequency\ of\ All\ Tripeptides}$$

Calculation of Conditional Probability of Middle Residue, Given it's N- and C- Terminal Residues

$$P(Y_M|X_N Z_C) = \frac{P_{(XYZ)}}{P_{(Y)}}$$

Calculation of Competency Score (CS-Score) of Middle Residue, Given N- and C- Terminal Residues

$$CS\text{-}Score\ (X_N Y_M Z_C) = 100\left(\frac{P(Y_M|X_N Z_C)}{P_{max}\ (AA_M|AA_N AA_C)}\right)$$

(B)

Consider Tripeptide, e.g. Lys-Ala-Met (K-A-M)

$$P\ (A) = \frac{Frequency\ (A)}{\sum Frequency\ of\ All\ Residues} = \frac{6.53E+05}{8.14E+06} = 8.03E-02$$

$$P\ (KAM) = \frac{Frequency\ (KAM)}{\sum Frequency\ of\ All\ Tripeptides} = \frac{1.04E+03}{8.06E+06} = 1.29E-04$$

$$P(A_M|K_N M_C) = \frac{P(KAM)}{P(A)} = \frac{1.29E-04}{8.03E-02} = 1.60E-03$$

$$P_{max}\ (AA_M|AA_N AA_C) = P(K_M|L_N E_C) = 1.09E-02$$

$$CS\text{-}Score\ (K_N A_M M_C) = 100\left(\frac{P(A_M|K_N M_C)}{P_{max}\ (K_M|L_N E_C)}\right) = 100\left(\frac{1.60E-03}{1.09E-02}\right)$$

$$CS\text{-}Score\ (K_N A_M M_C) = 14.08$$

(C)

Calculation of Probability of Individual Residue with a Specific SS and All Tripeptides with Specific SS from NPRD

$$P\ (Y^S) = \frac{Frequency\ (Y^S)}{\sum Frequency\ of\ All\ Residues}$$

$$P\ (X^S Y^S Z^S) = \frac{Frequency\ (X^S Y^S Z^S)}{\sum Frequency\ of\ All\ Tripeptides}$$

Calculation of Conditional Probability of Middle Residue with SS, Given it's N- and C- Terminal Residues with Specific SS Assignments

$$P(Y_M^S|X_N^S Z_C^S) = \frac{P(X^S Y^S Z^S)}{P(Y^S)}$$

Calculation of Competency Score (CSS-Score) of Middle Residue with SS, Given N- and C- Terminal Residues with Specific SS Assignments

$$CSS\text{-}Score\ (X_N^{S_x} Y_M^{S_y} Z_C^{S_z}) = 100\left(\frac{P\ \left(Y_M^{S_y}\middle|X_N^{S_x} Z_C^{S_z}\right)}{P_{max}\ \left(AA_M^{S_y}\middle|AA_N^{S_x} AA_C^{S_z}\right)}\right)$$

(D)

Consider Tripeptide, e.g. $\text{Lys}_{(H)}$-$\text{Ala}_{(H)}$-$\text{Met}_{(H)}$ (K_H-A_H-M_H)

$$P\ (A^H) = \frac{Frequency\ (A^H)}{\sum Frequency\ of\ All\ Residues} = \frac{3.55E+05}{8.14E+06} = 4.37E-02$$

$$P\ (K^H A^H M^H) = \frac{Frequency\ (K^H A^H M^H)}{\sum Frequency\ of\ All\ Tripeptides} = \frac{7.22E+02}{8.06E+06} = 8.96E-05$$

$$P(A_M^H|K_N^H M_C^H) = \frac{P(K^H A^H M^H)}{P\ (A^H)} = \frac{8.96E-05}{4.37E-02} = 2.05E-03$$

$$P_{max}\ (AA_M^H|AA_N^H AA_C^H) = P(M_M^H|E_N^H L_C^H) = 1.62E-02$$

$$CSS\text{-}Score\ (K_N^H A_M^H M_C^H) = 100\left(\frac{P\ (A_M^H|K_N^H M_C^H)}{P_{max}\ (M_M^H|E_N^H L_C^H)}\right) = \left(\frac{2.05E-03}{1.62E-02}\right)$$

$$CSS\text{-}Score\ (K_N^H A_M^H M_C^H) = 12.70$$

FIGURE 1 The overall flow of compiling scoring libraries. (A) A stepwise outline for calculating sequence-based competency score (CS-Score) of a residue by considering its adjacent residues toward N- and C-terminals. (B) A stepwise depiction for calculation of sequence-based competency score of an example tripeptide (Lys-Ala-Met is considered here). (C) A stepwise outline for calculating sequence and secondary structure-based competency score (CSS-Score) of a residue with a specific secondary structure by considering its adjacent residues toward N- and C-terminals with specific secondary structure assignment. (D) A stepwise depiction for calculation of sequence and secondary structure-based competency score of an example tripeptide (Lys(H)-Ala(H)-Met(H) is considered here)

where *X*, *Y*, and *Z* belong to any of the standard amino acid residues; S_x, S_y, and S_z belong to any of the three secondary structure assignments (H or E or C); $P\left(Y_M^{S_y}|Y_N^{S_x}Y_C^{S_z}\right)$ is the conditional probability of the middle residue "Y" having secondary structure "S_y", given a residue "X" having secondary structure "S_x" toward N-terminal and a residue "Z" having secondary structure "S_z" toward C-terminal; $P\left(X^{S_x}Y^{S_y}Z^{S_z}\right)$ is the probability of a tripeptide "XYZ" having secondary structure "$S_xS_yS_z$"; and $P\left(Y^{S_y}\right)$ is the probability of middle residue "Y" having secondary structure "S_y". It may be noted that "S" can assume any of the three secondary structure assignments (H, E, and C) but should be exactly the same for corresponding middle, N-terminal, and C-terminal residues to maintain the forward and backward neighborhood, and directionality of secondary structural triplets.

The conditional probabilities calculated in Equation (3) are further used for calculating sequence and secondary structure-based percentage competency score (CSS-Score) at an individual residue level by normalizing the conditional probabilities with the maximum conditional probability in all available combinations of the tripeptides having exactly same secondary structure assignment. The CSS-Score is calculated for the middle residue in a tripeptide with its secondary structure considering one adjacent residue of either side having specific secondary structure assignments as shown in Equation (4).

$$CSS-score\left(X_N^{S_x}Y_M^{S_y}Z_C^{S_z}\right) = 100\left(\frac{P\left(Y_M^{S_y}|X_N^{S_x}Z_C^{S_z}\right)}{P_{max}\left(AA_M^{S_y}|AA_N^{S_x}AA_C^{S_z}\right)}\right), \qquad (4)$$

where $CSS-score\left(X_N^{S_x}Y_M^{S_y}Z_C^{S_z}\right)$ is the sequence and secondary structure-based competency score of the middle residue "Y" having a secondary structure "S_y", given a residue "X" having a secondary structure "S_x" toward N-terminal and a residue "Z" having a secondary structure "S_z" toward C-terminal; $P\left(Y_M^{S_y}|X_N^{S_x}Z_C^{S_z}\right)$ is the conditional probability of the middle residue "Y" having a secondary structure "S_y", given a residue "X" having a secondary structure "S_x" toward N-terminal and a residue "Z" having a secondary structure "S_z" toward C-terminal; and $P_{max}\left(AA_M^{S_y}|AA_N^{S_x}AA_C^{S_z}\right)$ is the maximum conditional probability observed for any of the tripeptides with exactly the same secondary structure for middle, N-terminal, and C-terminal residues. The overall flow of computation of CSS-Scores is depicted in Figure 1C,D with an example tripeptide, Lys-Ala-Met with helices as secondary structure assignment for all the three residues. These scores are used to evaluate the overall competence of natural and nonnatural protein sequences. It is worth mentioning that the CS-Scores and CSS-Scores are nonzero and positive values which may be maximum up to 100.

The presently used 100% sequence identity filter ensured inclusion of the maximum number of possible triplets of amino acid residues and corresponding secondary structural elements. The normalization used in the Equations (2) and (4) calculates the score as a ratio of probabilities and removes the statistical bias due to closely related sequences. To investigate it further, all the natural protein sequences are clustered at lower sequence identity filters viz. 80%, 60%, and 40% and CSS-Scores libraries are compiled using Equations (3) and (4). A very high correlation is observed among the CSS-Scores of the triplets derived from the natural protein sequences at different sequence identity filters (40% and 100% = 0.95, 60% and 100% = 0.96%, 80% and 100% = 0.96) as shown in Figure S2. Considering the high similarity in CSS-Score libraries and the decline in triplet combinations at different sequence identity filters, it may be posited that the filtering at 100% sequence identity should not impart any bias to the statistics while ensuring the maximum utilization of available information at triplet level.

2.6 | Calculation of competency score for a protein sequence

For calculation of overall competency scores of a given protein sequence, the CS- and CSS-Scores of individual residues are used. It may be noted that the first residue (N-terminal residue) and the last residue (C-terminal residue) do not have their individual CS- and CSS-Scores. The overall CS- and CSS-Scores for a given protein may be calculated as shown in Equations (5) and (6).

$$Overall\,CS-Score_{Protein}=\frac{\sum_{i=2}^{i=(N-1)}CS-Score(i)}{(N-2)}, \tag{5}$$

where *N* is sequence length of the protein for which the overall CS-Score is to be calculated, *CS-Score(i)* is CS-Score of individual residues as calculated in Equation (2).

$$Overall\,CSS-Score_{Protein}=\frac{\sum_{i=2}^{i=(N-1)}CSS-Score(i)}{(N-2)}, \tag{6}$$

where *N* is sequence length of the protein for which the overall CSS-Score is to be calculated, *CSS-Score(i)* is CSS-Score of individual residues as calculated in Equation (4).

2.7 | Competency scores for natural proteins

The sequence and sequence and secondary structure scoring libraries (CS-scores and CSS-Scores) are used to calculate overall competency scores for individual sequences in natural proteins reference dataset of 41 132 proteins. The distribution curves for average competency scores in terms of CS-Scores and CSS-Scores are shown in Figure 2 (colored in green). Additionally, the scatter plots of average competency scores are shown in the Figure S4 for better insight. It is observed that the sequence-based competency scores (CS-Scores) averaged at 33.2 ± 3.14 and the sequence and secondary structure-based competency scores (CSS-Scores) averaged at 18.0 ± 3.61 for the reference dataset of natural protein sequences.

2.8 | Competency scores for nonnatural proteins (NNP-NC)

For all the computationally generated protein sequences in nonnatural protein (NNP-NC) reference dataset, the overall competency scores for individual sequences are calculated by using tripeptide-based CS-Scores and CSS-Scores. The distribution curves of CS-Scores and CSS-Scores for nonnatural protein (NNP-NC) sequences are shown in

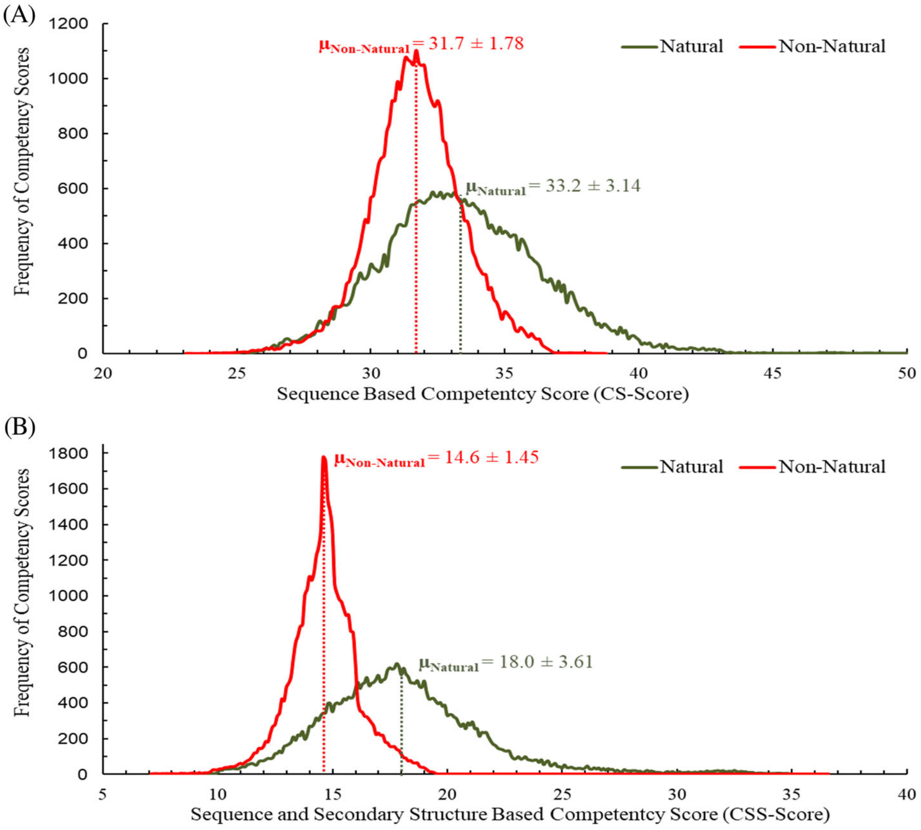


FIGURE 2 The distribution curves of competency scores. (A) Distribution curves for sequence-based competency scores of natural (in green color) and nonnatural (in red color) (CS-Scores). (B) Distribution curve for sequence and secondary structure-based competency scores (CSS-Scores) for natural (in green color) and nonnatural (in red color) proteins in corresponding reference datasets. The CSS-Scores are reflecting a better differentiation of natural and nonnatural proteins as compared to CS-Scores

Figure 2 (colored in red). Also, the scatter plots of competency scores of individual proteins in nonnatural proteins (NNP-NC) reference dataset are depicted in the Figure S5. In case of reference dataset of nonnatural protein (NNP-NC) sequences, the observed sequence-based competency scores averaged at 31.7 ± 1.78 and the sequence and secondary structure-based competency scores averaged at 14.6 ± 1.45.

In the present study, the secondary structure prediction is used to estimate the likelihood of secondary structure for the nonnatural proteins which is further utilized in performing overall scoring of nonnatural proteins. Further, the method used here for the secondary structure prediction is not exclusively dependent on amino acid substitution matrix (viz. BLOSUM62), it also implements three different neural network weights which are expected to improve the prediction accuracy.

2.9 | Differences in competency scores of natural proteins (NP) and nonnatural proteins (NNP-NC)

It is very difficult to conclude directly from the average competency scores for natural proteins (NP) and nonnatural protein sequences (NNP-NC) if these deviates meaningfully. For testing the significance of differences in the average competency scores for natural (NP) and nonnatural protein (NNP-NC) sequences in the reference datasets, *z*-test of two samples for means is conducted on competency scores of 41 132 natural protein sequences and 41 132 nonnatural protein sequences (41 132 observations each). Based on the outcome of *z*-test, in case of sequence-based competency scores (CS-Scores), the natural protein sequences (μ = 33.2, σ = 3.14, n = 41 132) and nonnatural protein (NNP-NC) sequences (μ = 31.7, σ = 1.78, n = 41 132) are hypothesized to be different. The difference is very significant, *z* = 96.73, *P* = .00 (two-tail). Also, in case of sequence and secondary structure-based competency scores (CSS-Scores), the natural protein sequences (μ = 18.0, σ = 3.61, n = 41 132) and nonnatural protein (NNP-NC) sequences (μ = 14.6, σ = 1.45, n = 41 132) are hypothesized to be different. The difference is very significant, *z* = 210.75, *P* = .00 (two-tail). Further details of *z*-test statistics are provided in the Table S1.

To investigate the differences in amino acid neighbor preferences in different secondary structures for computationally generated nonnatural protein sequences (NNP-NC), we derived the conditional probability of triplets using these sequences and corresponding predicted secondary structures. The conditional probabilities of triplets in natural proteins and computationally generated nonnatural proteins showed a correlation of 0.73, which supports the assumption that the computationally generated protein sequences have differences in amino acid neighbor preferences in different secondary structures. These differences in the neighboring preferences may help in computational sampling of protein sequences with higher potential of mimicking the natural proteins. Further, to investigate the possibility of computational bias, instead of their original secondary structures, the predicted secondary structures for the natural proteins are used to recalculate the CSS-Score libraries. It is observed that the CSS-Score libraries computed using predicted secondary structures showed a significant similarity (*r* = 0.94) with the CSS-Score libraries computed using the experimental secondary structures. Additionally, a *z*-test is performed to further analyze the differences in the CSS-Score libraries as reported in the supplementary materials (Table S2). The CSS-Score library derived from experimental secondary structures of natural protein (μ = 5.68, σ = 8.00, n = 91 222) and the CSS-Score library derived from predicted secondary structure of natural proteins (μ = 5.69, σ = 8.29, n = 91 222) are hypothesized to be significantly similar (*z* = –0.39, *P* = .69 (two-tail)). As the difference is not significant, it may be assumed that in case of sampling and scoring novel proteins, the performance of the proposed scoring function may not change significantly upon using the predicted secondary structures.

2.10 | Efficacy of competency scores

The receiver operating characteristic curve (ROC Curve) is one of the most prevalent and extensively instigated statistical tools for assessing the discriminatory efficacy of a given classifier. Here, the average competency scores of the individual proteins at sequence (CS-Score) and sequence and secondary structural (CSS-Score) levels are assessed for their potential to differentiate natural (NP) and nonnatural protein (NNP-NC) sequences. Under the assumption that the higher CS-Score and CSS-Score for a protein are indicative of its imitating behavior as natural proteins and lower CS-Score and CSS-Score for a protein are suggestive of imitating behavior as nonnatural proteins (NNP-NC). At different threshold values of competency scores, different pairs of sensitivity and specificity are derived from reference dataset of natural and nonnatural proteins using Equation (7) as follows.

$$ROC\,(t) = \{(FPR(t), TPR(t)), t \in (Range\ of\ Competency\ Score)\}, \qquad (7)$$

where *FPR*(*t*) is the false positive rate at a threshold value "*t*"; *TPR*(*t*) is the true positive rate at a threshold value "*t*".

The ROC curves are plotted with two underlying assumption, (a) the potential of competency scores to identify the natural proteins (NP) and (b) potential of competency scores to identify the nonnatural proteins (NNP-NC). At different thresholds of CS-Scores, $\mathrm{ROC}(t)_{\mathrm{Natural}}$, and $\mathrm{ROC}(t)_{\mathrm{Nonnatural}}$ are calculated and plotted in Figure 3A. Similarly, at different thresholds of CSS-Scores, $\mathrm{ROC}(t)_{\mathrm{Natural}}$, and $\mathrm{ROC}(t)_{\mathrm{Nonnatural}}$ are calculated and plotted in Figure 3B. The threshold values at point of intersection ROC curves of natural and nonnatural proteins are observed to be the optimum cutoff for differentiating natural and nonnatural proteins.

The ROC curves of CS-Score for natural proteins (NP) and nonnatural protein (NNP-NC) sequences are intersecting at a threshold value of 32.15. At intersection point, the sensitivity and specificity in identifying natural (NP) and nonnatural proteins (NNP-NC) is 0.62 and 0.63, respectively. However, the Mathews Correlation Coefficient (MCC) at CS-Score cutoff value of 32.15 is 0.26 which indicates weak prediction model for binary classification. The low value of MCC is

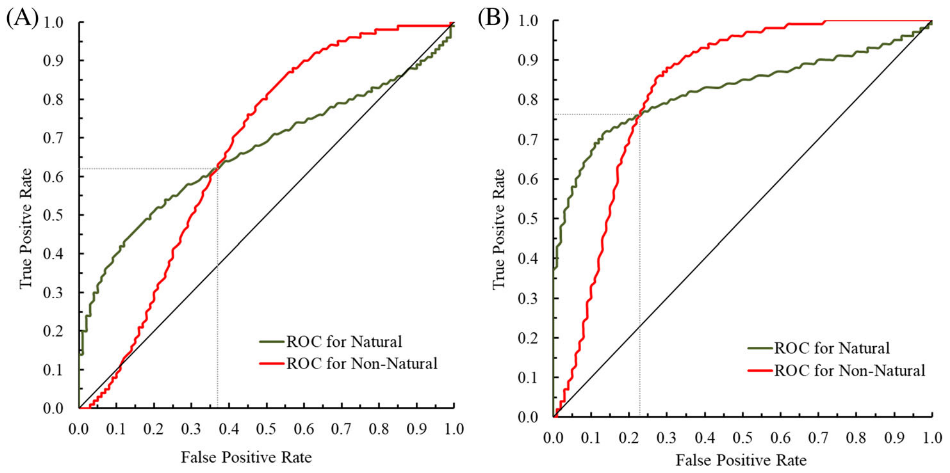


FIGURE 3 The ROC curves for identifying natural and nonnatural proteins. (A) CS-Scores threshold-based ROC curves for natural and nonnatural proteins. (B) CSS-Scores threshold-based ROC curves for natural and nonnatural proteins [Color figure can be viewed at wileyonlinelibrary.com]

suggestive of inability of CS-Score in discriminating natural and nonnatural proteins. The ROC curves of CSS-Score for natural (NP) and nonnatural proteins (NNP-NC) showed a considerably improved sensitivity and specificity at their intersection threshold value. The CSS-Score based ROC curves intersected at 15.5 where the sensitivity is 0.76 and specificity is 0.77. Notably, the Mathews Correlation Coefficient (MCC) at CSS-Score cutoff value of 15.5 is 0.54 which is suggestive of strong prediction model for binary classification. The calculation of sensitivity, specificity, and Mathews Correlation Coefficient is explained in supplementary information (Supplementary Note II). From ROC curves, sensitivity, specificity, and MCC values, it may be interpreted that the only sequence-based competence score (CS-Score) is not very efficient in discriminating natural (NP) and nonnatural proteins (NNP-NC). However, the sequence and secondary structure-based competency score (CSS-Score) reflects a promising potential of discriminating natural (NP) and nonnatural proteins (NNP-NC). The performances of CS- and CSS-Score are further evaluated on different datasets and discussed in results section.

2.11 | Competency score analysis at residue level in individual sequences

The efficacy of competency scores in discriminating natural proteins (NP) and nonnatural protein (NNP-NC) sequences does not elucidate the extent of its prediction reliability. To explore this further, a residue level analysis of competency scores of individual proteins of natural and nonnatural reference datasets is performed. In case of CS-Scores, if a protein sequence is classified as natural protein (NP) on the basis of its overall competency score (CS-Score ≥ 32.15) and more than 59% of its residues are scoring above the threshold, then it is scoring better than 80% of the natural proteins in reference dataset and may be considered as natural protein with 80% confidence value. The required number of percentage residues scoring above the threshold in a protein increases to 69% for it to score better than 95% of the natural proteins in reference dataset. Likewise, if a protein is classified as nonnatural protein on the basis of competency score (CS-Score < 32.15), and more than 61% of its residues are scoring below the threshold, then it is scoring better than 80% of the nonnatural proteins and qualifies as nonnatural protein with 80% confidence value. The required number of percentage residues scoring below the threshold in a protein increases to 67% for it to be classified as nonnatural protein with 95% possibility.

In case of CSS-Scores, for classifying a protein as natural protein, having scored better than 80% of natural proteins in reference dataset, it needs to have more than 62% of its residues scoring above the threshold (CSS-Score ≥ 15.50). The required percentage number of residues scoring above the threshold increases to 72% for classifying a protein as natural with score better than 95% natural proteins. For identifying a protein as nonnatural protein having outscored 80% of nonnatural proteins, 64% of its residues must be scoring below the threshold (CSS-Score < 15.50). The percentage number of residues scoring below the threshold increases to 70% for identifying a protein as nonnatural with outscoring 95% of nonnatural proteins. Figure 4A shows the distribution of percentage number of proteins in natural and nonnatural proteins reference datasets (on *y*-axis) against different cutoffs of percentage residues scoring below the threshold values of CS-Scores. Similarly, Figure 4B shows the distribution of percentage number of proteins in natural and nonnatural proteins reference datasets against different cutoffs of percentage residues scoring below the threshold values of CSS-Scores. It is worth mentioning that in case of natural proteins, the percentage of residues above threshold is considered while in case of nonnatural proteins, the percentage of residues scoring below the threshold is accounted. Thus, in case of natural protein while referring to Figure 4, the percentage number of proteins at different values of percentage number of residues scoring above the threshold can be calculated by subtracting the corresponding value from 100. Additionally, the different values of percentage residues scoring above and below the threshold for natural and nonnatural proteins in reference datasets are furnished in supplementary Table S3. It may be noted that the extent of overlap in percentage number of residues cutoffs is relatively less in case of CSS-Scores which is indicative of its better discriminating potential of natural and nonnatural proteins. The same is demonstrated on the different datasets and discussed in results section.

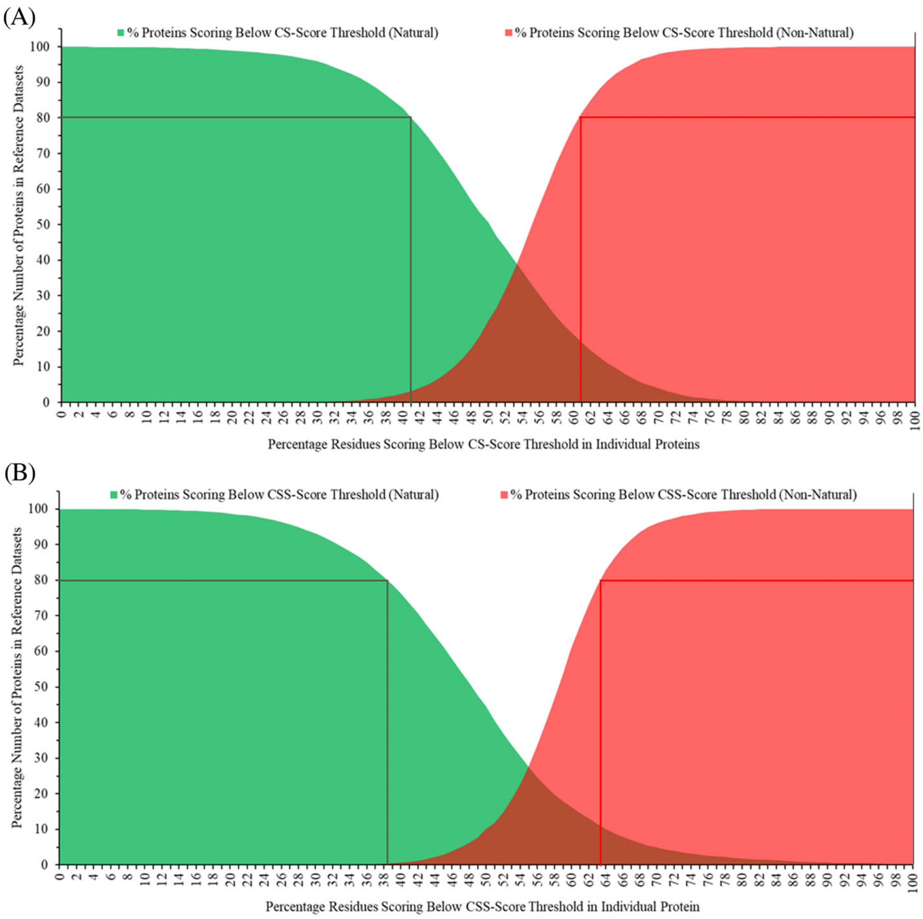


FIGURE 4 Distribution of natural (in green) and nonnatural (in red) proteins at different values of their percentage residues scoring below the derived cutoffs from ROC curves. (A) CS-Score based distribution, highlighting percentage residues scoring below threshold for 80% of natural and nonnatural proteins. (B) CSS-Score based distribution, highlighting percentage residues scoring below threshold for 80% of natural and nonnatural proteins [Color figure can be viewed at wileyonlinelibrary.com]

2.12 | Competency score based prediction of example protein

For a given target protein, the sequence-based competency scores (CS-Score) for each residue (except for first and last residues) are calculated using precompiled CS-Scores libraries (explained in Section 2.4). The overall competency score is calculated from the scores of individual residues as shown in Equation (5). Based on the average CS-Score, the protein is predicted as natural (CS-Score ≥ 32.15) or nonnatural protein (CS-Score < 32.15). Also, the percentage of residues scoring below or above the threshold are calculated and employed for deriving the possibility of the predictions accuracy by comparing it with the values to the distribution of natural and nonnatural proteins. The overall flow of carrying out CS-Score based prediction of a target protein is demonstrated in Figure 5A. Further, the secondary structure prediction of a target protein (if not known) is performed using the standalone version of PSIPRED. The sequence and secondary structure information is applied for calculating sequence and secondary structure-based competency scores (CSS-Scores) for each residue (except for first and last residues) by utilizing the precompiled CSS-Scores libraries. The overall CSS-Score for the target protein is calculated and used for classifying it as natural (CSS-Score ≥ 15.50) or nonnatural (CSS-Score < 15.50). The percentage of residues scoring above or below the threshold is used for deriving the possibility of prediction accuracy via a comparison to the distribution of natural and nonnatural proteins in reference datasets. The overall flow of performing CSS-Score based prediction of a target protein is demonstrated in Figure 5B. In case of CS-Score based prediction (Figure 5A), the example target protein is identified as natural protein (CS-Score ≥ 32.15), having about 39% residues scoring below threshold (61.2% residues scoring above threshold). Referring to Table S3 (column 1 and 2, row 39), the example target protein is scoring better than 85% natural proteins. Likewise, in case of CSS-Score based prediction (Figure 5B), the example target protein is identified as natural protein (CSS-Score ≥ 15.50), having about 30% residues scoring below threshold (69.9% residues scoring above threshold). Referring to Table S3 (column 1 and 4, row 30), the example target protein is scoring better than 93% natural proteins.

Since the competency score libraries and threshold values are precomputed from reference datasets of natural and nonnatural proteins, the batch calculation of CS-Score and CSS-Score is very time and computationally efficient.

3 | RESULTS AND DISCUSSION

The performance of CS- and CSS-Scores is evaluated on a test dataset of natural proteins (NP) and nonnatural proteins (NNP-NC) (17 626 proteins each, as mentioned in Section 2.3). Additionally, a dataset of

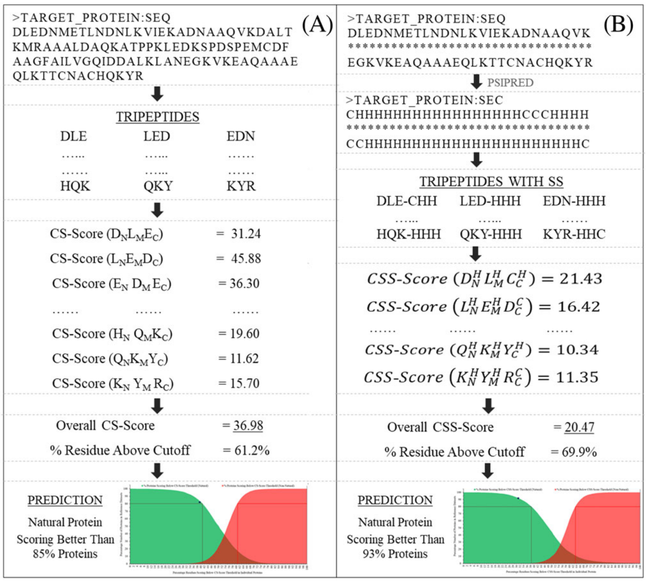

(A)
>TARGET_PROTEIN:SEQ
DLEDNMETLNDNLKVIEKADNAAQVKDALT
KMRAAALDAQKATPPKLEDKSPDSPEMCDF
AAGFAILVGQIDDALKLANEGKVKEAQAAAE
QLKTTCNACHQKYR
TRIPEPTIDES
DLE LED EDN
HQK QKY KYR
CS-Score (D_N L_M E_C) = 31.24
CS-Score (L_N E_M D_C) = 45.88
CS-Score (E_N D_M E_C) = 36.30
CS-Score (H_N Q_M K_C) = 19.60
CS-Score (Q_N K_M Y_C) = 11.62
CS-Score (K_N Y_M R_C) = 15.70
Overall CS-Score = 36.98
% Residue Above Cutoff = 61.2%
PREDICTION
Natural Protein Scoring Better Than 85% Proteins
(B)
>TARGET_PROTEIN:SEQ
DLEDNMETLNDNLKVIEKADNAAQVK
EGKVKEAQAAAEQLKTTCNACHQKYR
PSIPRED
>TARGET_PROTEIN:SEC
CHHHHHHHHHHHHHHHHHHHCCCHHHH
CCHHHHHHHHHHHHHHHHHHHHHHHHC
TRIPEPTIDES WITH SS
DLE-CHH LED-HHH EDN-HHH
HQK-HHH QKY-HHH KYR-HHC
CSS-Score (D^H_N L^H_M C^H_C) = 21.43
CSS-Score (L^H_N E^H_M D^C_C) = 16.42
CSS-Score (Q^H_N K^H_M Y^H_C) = 10.34
CSS-Score (K^H_N Y^H_M R^C_C) = 11.35
Overall CSS-Score = 20.47
% Residue Above Cutoff = 69.9%
PREDICTION
Natural Protein Scoring Better Than 93% Proteins

FIGURE 5 Demonstration for prediction of a target protein as natural or nonnatural protein. (A) Prediction for target protein derived from average CS-Score and the percentage of residues scoring above the threshold. (B) Prediction for a target protein derived from average CSS-Score and the percentage of residues scoring above the threshold [Color figure can be viewed at wileyonlinelibrary.com]

~57 000 unique natural proteins (clustered at 40% sequence identity) of sequence length varying from 50 to 700 residues from UniProtKB is selected after excluding all the natural proteins of SCOPe database (58 758 proteins). Further, two more datasets of ~57 000 computationally generated proteins, one with natural AA compositions and another with uniform amino acid compositions constraint (NNP-NC and NNP-UC) are considered for quantifying the ability of competency scores in differentiating natural proteins from nonnatural proteins.

3.1 | Evaluation on natural and nonnatural proteins in test datasets

The test datasets of natural proteins (NP) and nonnatural proteins (NNP-NC) are subjected to calculation of competency scores. For CSS-Score calculation, the secondary structure of individual natural protein is extracted from the corresponding structure, while the secondary structure of individual nonnatural protein (NNP-NC) is predicted using PSIPRED. The overall CS- and CSS-Scores of proteins in test datasets are calculated and further used for categorizing them into natural and nonnatural based on the threshold values (Natural Proteins ≥ CS-Score 32.15 > Nonnatural Proteins; (Natural Proteins ≥ CSS-Score 15.50 > Nonnatural Proteins). The evaluation statistics of CS- and CS-Scores are reported in Table 1. The distribution of CS- and CSS-Scores for all the proteins in the Test Dataset is shown in Figure S6. Here, it may be noted that the CSS-Score based categorization of natural and nonnatural proteins outperformed CS-Score based categorization. It reflects the gain in prediction accuracy with the addition of secondary structure information.

TABLE 1 Assessment of CS- and CSS-scores on test datasets of 35 252 proteins for identifying natural and nonnatural proteins

Statistics	CS-score	CSS-scores
Sensitivity	0.62	0.76
Specificity	0.62	0.75
Accuracy	0.62	0.74
Mathews correlation coefficient	0.25	0.53

3.2 | Evaluation on external dataset of natural and nonnatural sequences

For assessing the performance of the proposed competency scores, an independent dataset of reviewed proteins from UniProtKB is extracted by filtering out all the natural proteins considered in reference and test datasets of natural proteins. The filtered reviewed proteins are further clustered to 40% sequence identity to eliminate the closely related proteins which resulted into 56 637 proteins. This dataset of unique reviewed proteins from UniProtKB is referred to as external dataset of natural proteins. For all the proteins in external dataset of natural proteins, the CS- and CSS-Scores are calculated and compared to threshold values identified in methods section, that is, Natural Proteins (CS-Score ≥ 32.15; CSS-Score ≥ 15.50). Based on CS-Score threshold, it is observed that only 33 729 (59%) proteins could be identified as natural proteins. However, CSS-Score based evaluation performed much better by identifying 45 876 (81%) proteins as natural proteins.

In the nonnatural proteins dataset of 58 758 proteins, the amino acid compositions were constrained to corresponding amino acid compositions of natural proteins. Similarly, another dataset of nonnatural proteins comprising 60 000 proteins is computationally generated and clustered at 40% sequence identity to ensure the absence of similar proteins. Further, the clustered proteins are screened against the previously considered nonnatural proteins dataset to filter out the similar proteins. The clustering and filtering resulted in a new dataset of nonnatural proteins comprising 56 873 unique proteins, entirely independent of the nonnatural proteins used in deriving thresholds for CS- and CSS-Scores. This new dataset of 56 836 nonnatural proteins is referred to as external dataset of nonnatural proteins (NNP-NC). The CS- and CSS-Scores for the external dataset of nonnatural proteins are calculated and classified using the previously derived thresholds for nonnatural proteins (CS-Score < 32.15; CSS-Score < 15.50). The CS-Score based identification of nonnatural proteins categorized 48 324 (85%) proteins as nonnatural while CSS-Score could identify 51 153 (90%) proteins as nonnatural.

So far in this study, we have used natural proteins, adopted from SCOPe and UniProtKB databases, and nonnatural proteins, computationally generated with the same amino acid composition as the natural proteins in UniProt database. Further, in this study, a dataset of computationally generated 60 000 random proteins, with all amino acid residues having equal probability of occurrence, is used for evaluating the potential of the competency scores in discriminating natural proteins from randomly generated proteins. This dataset of nonnatural proteins with uniform composition of amino acid residues (NNP-UC) is clustered at 40% sequence identity which resulted in 57 374 unique nonnatural proteins. All these proteins scored within the threshold derived for nonnatural proteins (CS-Score < 32.15; CSS-Score < 15.50). A distribution of CS- and CSS-Scores of the proteins in the external dataset of nonnatural proteins is shown in Figure 6. It is worth mentioning that the CSS-Score based categorization of the natural, nonnatural, and the random proteins performed consistently on a considerably higher side. Clearly, the CSS-Score emerges as a far better measure than CS-Score for identifying natural proteins from nonnatural and random proteins. A summary of performance of CS- and CSS-Score in identifying natural and nonnatural protein sequences in external dataset of proteins is provided in Table 2.

For further statistical evaluation, the external datasets are combined where the natural proteins are tagged as positives, and nonnatural and random proteins are tagged as negatives. For CS-Score identification, the sensitivity, specificity, and Mathews correlation coefficient are observed to be 0.60, 0.92, and 0.57, respectively. Likewise, for CSS-Score based identification, the sensitivity, specificity, and Mathews correlation coefficient are 0.81, 0.95, and 0.79, respectively.

3.3 | Benchmarking with existing methods

The CS- and CSS-Score based identification of natural and nonnatural proteins is further benchmarked with existing methods. It is worth noting that there are not many methods available for directly scoring the protein sequences to classify them as natural and nonnatural proteins. Here, we benchmarked the present scoring method with FoldIndex[40] and FOLD.[20,41] The FoldIndex method implements average residue hydrophobicity and net charge to derive the foldability or unfoldability of a given protein sequence where the positive score represents foldable and the negative score represents unfoldable. Some of the proteins scored very close to zero (−0.005 ≤ SCORE ≤ 0.005) which are accounted as unreliable prediction. A very recently proposed method, named FOLD, utilizes the precomputed triplet (FOLD3) and quadruplet (FOLD4) frequencies in natural and random protein sequences to classify a given protein sequence into any of the four classes, viz. sure

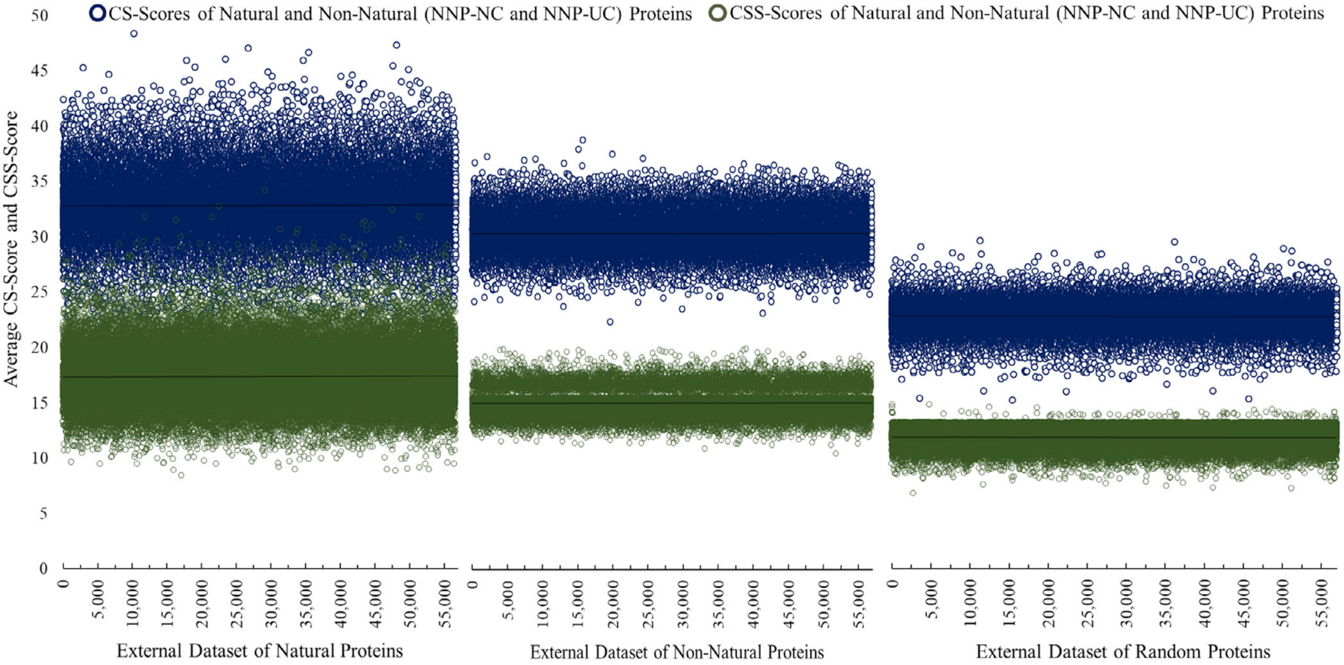


FIGURE 6 A comparison of CS- and CSS-Scores across external datasets of natural, computationally generated nonnatural (NNP-NC and NNP-UC). A downward trend is observed from natural proteins to nonnatural proteins [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 A summary of CS- and CSS-score identification on different external datasets

External dataset	Total number of proteins	CS-score natural	CS-score nonnatural	CSS-score natural	CSS-score nonnatural
Natural	56 637	33 729	22 908	45 876	10 761
Nonnatural (NNP-NC)	56 873	8549	48 324	5720	51 153
Nonnatural (NNP-UC)	57 374	0	57 374	0	57 374

folded, sure random, guessed folded, and guessed random. While benchmarking our method, we combined the sure folded and guessed folded as the natural proteins, and the sure random and guessed random as the nonnatural proteins. The summary of the predictions using FoldIndex, FOLD, CS-Score, and CSS-Score for external datasets of natural, nonnatural, and random proteins is shown in Table 3 and Figure S8. For calculating sensitivity and specificity, the protein scored as unreliable prediction are not considered.

Some other methods developed for characterization of protein sequences into natural and random proteins,[4-6,29,42] could not be independently validated on the dataset of 170 884 proteins due to unavailability of standalone versions. For such methods, the evaluation statistics reported in respective research article is compiled and provided in Table 4.

The benchmarking of CS- and CSS-Score with previously reported methods in Table 3 demonstrates a reasonably better performance in terms of sensitivity, specificity, and accuracy. Despite the fact that the accuracy of the methods accounted in Table 4 is adopted from respective research article, which is only restricted to a small dataset of natural and random proteins in most cases, the accuracy of CSS-Score clearly outperformed most of these methods except Lucrezia et al[4] which is validated on a dataset of 1500 small proteins of ∼70

TABLE 3 The summary of benchmarking of CS-score and CSS-score with FoldIndex and FOLD on the dataset of 170 884 proteins, comprising 56 637 natural and 114 247 nonnatural proteins

Method	Predicted natural	Predicted nonnatural	Unreliable prediction	Sensitivity	Specificity	Percentage accuracy (%)
FoldIndex	140 767	16 815	13 302	0.86	0.10	35
FOLD3*	63 941	105 924	1019	0.61	0.74	69
FOLD4*	49 443	107 508	13 933	0.63	0.84	77
FOLD5*	40 719	121 801	8364	0.41	0.82	69
CS-Score	42 278	128 606	0	0.60	0.92	82
CSS-Score	51 596	119 288	0	0.81	0.95	91

TABLE 4 Summary of articles reporting characterization of natural and random proteins by implementing various approaches

Method/reference	Parameters/approach	Dataset (N + R)	Accuracy (%)	Remark

Munteanu et al, 2008	Star network topological indices	N = 1046
		R = 1046
			90	Bias for random

Santoni et al, 2016	ML on proximity measure between pair of amino acids	N = 1047
		R = 10 470
			75	Small dataset for natural

Garbuzynskiy et al, 2004	Hydrophobicity and contact number	N = 80
		R = 90
			83	Small dataset for natural

De Lucrezia et al, 2012	Evolutionary neural network on small protein (∼70 aa)	N = 762
		R = 762
			94	Only small proteins accounted

Tsygvintsev, 2019	Neural network based on time series analysis	N = 3502
		R = 3502
			85	24D vector used in complex training

Present study CS-score	Competency Scores derived from sequences	N = 56 636
		R = 114 247
			82	Relatively lower accuracy

Present study CSS-score	Scores derived from sequences and 2° structures	N = 56 636
		R = 114 247
			91	

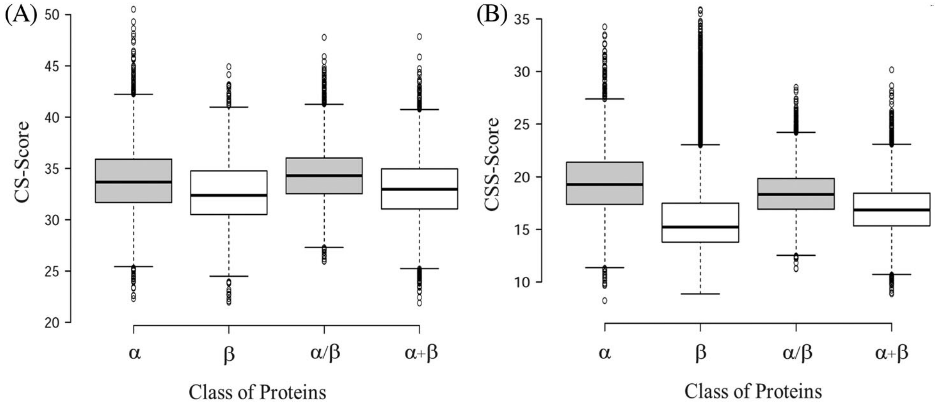


FIGURE 7 A boxplot representation of (A) CS- and (B) CSS-Scores for four classes of proteins (α, β, α/β, and α + β)

TABLE 5 Summary of compiled designed proteins from previous research articles

Research article	Designed proteins	Expressed in *E. coli*	Reported soluble	Monomeric proteins	Solved structures
Koga et al, 2012	54	49	45	19	16
Lin et al, 2015	49	49	45	31	10
Koepnick et al, 2019	144	119	99	65	55
Total	247	217	189	115	81

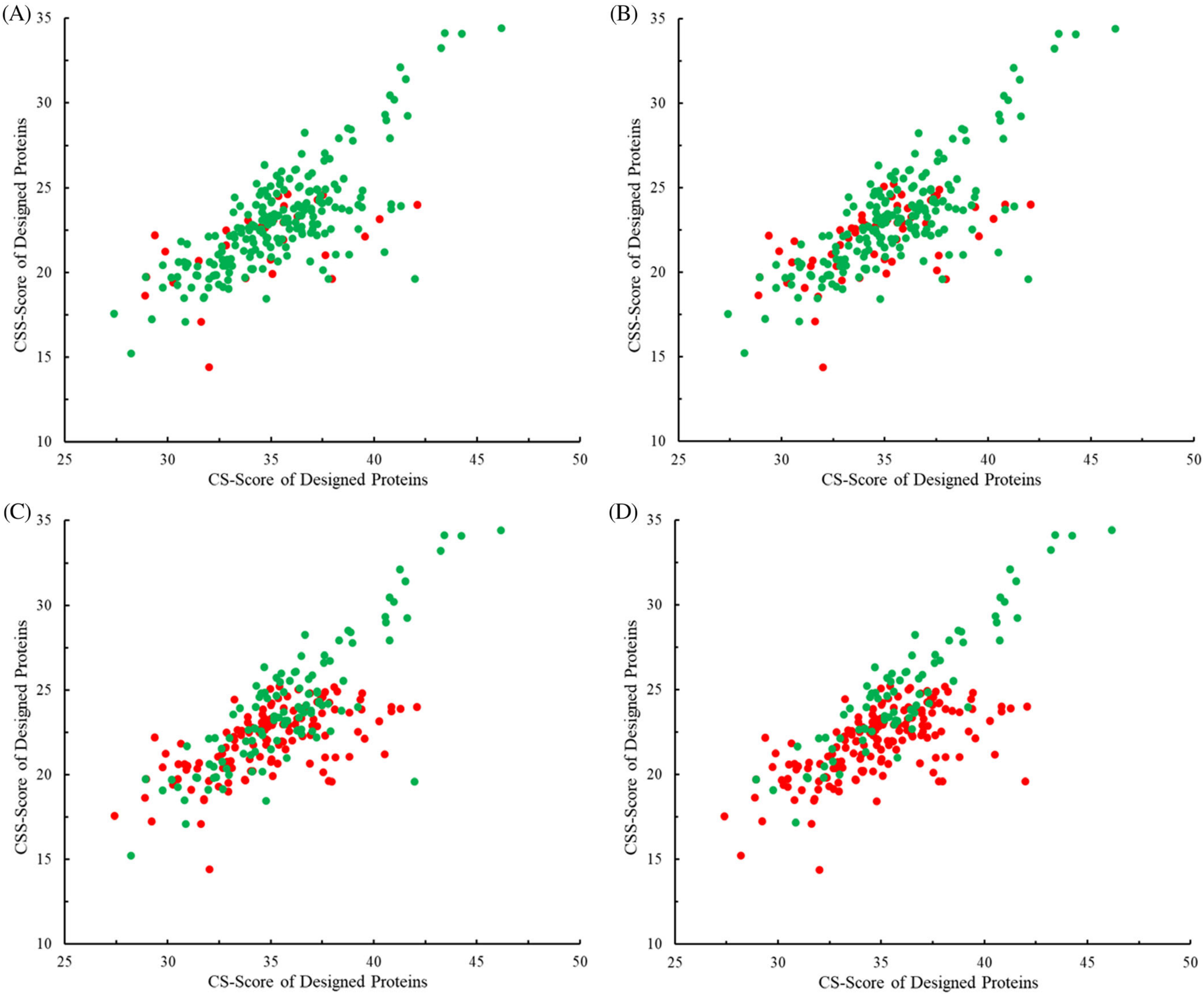


FIGURE 8 Competency score-based analysis of successful (green) designed and failed (red) proteins at (A) expression level, (B) solubility level, (C) oligo-state level, and (D) structural level [Color figure can be viewed at wileyonlinelibrary.com]

amino acid residues length. Also, the methods reported by Munteanu et. al (26) was cross validated by Santoni et al[6] to report an accuracy of 79% with a very low true positive rate.

3.4 | Distribution for different protein classes

The competency scores are calculated for the unique protein sequences of all alpha (α), all beta (β), alpha and beta (α/β), and alpha plus beta (α + β) proteins representing 289, 178, 148, and 388 protein folds. The average CS-Scores are observed to be 33.9 (±3.40), 32.7 (±3.00), 34.4 (±2.67), and 33.1 (±3.06) for all alpha (α), all beta (β), alpha and beta (α/β), and alpha plus beta (α + β) proteins, respectively. Likewise, the average CSS-Scores are found to be 19.5 (±3.20), 16.5 (±4.25), 18.4 (±2.27), and 16.9 (±2.50) for all alpha (α), all beta (β), alpha and beta (α/β), and alpha plus beta (α + β) proteins, respectively. A boxplot representation of CS- and CSS-Scores is shown in Figure 7 and the additional statistics are provided in Table S4.

It is worth noting that in case of all protein classes the average CS- and CSS-Scores are beyond the minimum threshold for natural proteins that is, CS-Score ≥ 32.15 and CSS-Score ≥ 15.50, respectively. However, a further investigation is required to find out if the scores are significantly deviating among different classes of proteins.

3.5 | Performance on reported designed proteins

A set of 247 designed protein sequences, reported in some previous research articles[43-45] is compiled for calculating the sequence and secondary structure-based competency scores. The experimental results of these designed proteins sequences are available for their expression, solubility, monomeric state, and structure. According to their respective articles, these sequences are selected for experimental validation after screening through some comprehensive scoring functions from more than 100 folds sampled sequences. Since only top ranked protein sequences (less than 0.1% of sampled sequences) are considered for experimental characterization, these are likely to score much higher than the expected competency scores of natural proteins. The details of the designed protein dataset are provided in Table S5 and a summary is provided in Table 5.

In total, 81 designed proteins could be solved as well characterized protein tertiary structures using X-ray crystallography and/or NMR methods. The rationale of screening the designed protein sequences using the CS- and CSS-Scores is to quantify the ability of these scores at expression, solubility, oligo-state, and structural level. In Figure 8, the CS- and CSS-Scores of designed proteins accounted in Table 5 are plotted as success (green circles) and failure (red circles) cases at expression, solubility, oligo-state, and structural levels.

It is observed that most of the proteins except one scored beyond the minimum threshold of natural proteins for CSS-Score (above 15.50). However, the same is not true for CS-Score as several proteins scored below the minimum threshold (below 32.15). It may also be noted that as we move from expression to solubility to oligo-state to structure, the upper-right quadrant (with CS-Score > 35 AND CSS-Score > 25) of the plots remains occupied by successful cases at all four levels. This observation may help in designing novel protein sequences with a higher potential of being successful at experimental validation.

4 | CONCLUSION

The infinitesimally small sequence space naturally scouted in the millions of years of evolution suggests that the natural proteins are impeded by some specific prerequisites and should diverge from computationally generated nonnatural protein sequences. Considering this, here we studied natural and computationally generated nonnatural proteins to develop a protein sequence fitness scoring function. The scoring function implements sequence and corresponding secondary structural information at tripeptide levels to differentiate natural and nonnatural proteins. The proposed scoring function is extensively validated on a dataset of about 210 000 natural and nonnatural protein sequences and benchmarked with existing methods for differentiating natural and nonnatural proteins. The high sensitivity, specificity, and percentage accuracy (0.81%, 0.95%, and 91% respectively) of the scoring function demonstrates its potential application for sampling the protein sequences with higher probability of mimicking natural proteins. Also, the four major classes of proteins (α proteins, β proteins, α/β proteins, and α + β proteins) are separately analyzed and β proteins are observed to scoring slightly on the lower side as compared to other classes. Further, an analysis of about 250 designed proteins (adopted from previously reported cases) helped in defining the boundaries for sampling the ideal protein sequences which may prove advantageous in computational protein design regimes.

ACKNOWLEDGMENTS

Authors are very grateful to Prof. Mihaly Mezei for help with running FOLD3 and FOLD4. We gratefully acknowledge the suggestions from Prof. Jaime Prilusky for automating the predictions using FoldIndex. We acknowledge RIKEN ACCC for the computing resource used in this study. This work is supported by Kakenhi 18H02395 from Japan Society for the Promotion of Science.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All the datasets used in the present study are provided at http://github.com/KYZ-LSB/ComProDes. Additionally, the programs for running the proposed protein sequence fitness scoring function, user tutorial, and readme files are provided for future use of the programs. There is no additional dependency required for running the programs in Linux environment.

ORCID

Kam Y. J. Zhang https://orcid.org/0000-0002-9282-8045

REFERENCES

1. Kc DB. Recent advances in sequence-based protein structure prediction. *Brief Bioinform*. 2017;18(6):1021-1032.
2. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science*. 2017;355(6322):294-298.
3. Trainor K, Broom A, Meiering EM. Exploring the relationships between protein sequence, structure and solubility. *Curr Opin Struct Biol*. 2017;42:136-146.
4. 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(5).e36634. http://dx.doi.org/10.1371/journal.pone.0036634.
5. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. To be folded or to be unfolded? *Protein Sci*. 2004;13(11):2871-2877.
6. 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.

7. Turjanski P, Ferreiro DU. On the natural structure of amino acid patterns in families of protein sequences. *J Phys Chem B*. 2018;122(49): 11295-11301.
8. Uversky VN. What does it mean to be natively unfolded? *Eur J Biochem*. 2002;269(1):2-12.
9. Lu PL, Min DY, DiMaio F, et al. Accurate computational design of multipass transmembrane proteins. *Science*. 2018;359(6379):1042-1046.
10. Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327.
11. Voet ARD, Noguchi H, Addy C, Zhang KYJ, Tame JRH. Biomineralization of a cadmium chloride nanocrystal by a designed symmetrical protein. *Angew Chem Int Edit*. 2015;54(34):9857-9860.
12. Brunette TJ, Parmeggiani F, Huang PS, et al. Exploring the repeat protein universe through computational protein design. *Nature*. 2015; 528(7583):580.
13. Voet ARD, Noguchi H, Addy C, et al. Computational design of a self-assembling symmetrical beta-propeller protein. *Proc Natl Acad Sci U S A*. 2014;111(42):15102-15107.
14. Burke AJ, Lovelock SL, Frese A, et al. Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature*. 2019;570(7760):219.
15. Langan RA, Boyken SE, Ng AH, et al. De novo design of bioactive protein switches. *Nature*. 2019;572(7768):205.
16. Wang TT, Fan XT, Hou CX, Liu JQ. Design of artificial enzymes by supramolecular strategies. *Curr Opin Struct Biol*. 2018;51:19-27.
17. Welborn VV, Head-Gordon T. Computational design of synthetic enzymes. *Chem Rev*. 2019;119(11):6613-6630.
18. Leelananda SP, Jernigan RL. Diversity of sequences folding to highly and poorly designable structures. *Biophys J*. 2012;102(3):456.
19. Tian PF, Best RB. How many protein sequences fold to a given structure? A coevolutionary analysis. *Biophys J*. 2017;113(8):1719-1730.
20. Mezei M. On predicting foldability of a protein from its sequence. *Proteins*. 2019;88(2):355–365.
21. Laurenzi A, Hung LH, Samudrala R. Structure prediction of partial-length protein sequences. *Int J Mol Sci*. 2013;14(7):14892-14907.
22. LaBean TH, Butt TR, Kauffman SA, Schultes EA. Protein folding absent selection. *Genes*. 2011;2(3):608-626.
23. Angyan AF, Perczel A, Gaspari Z. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Lett*. 2012;586(16):2468-2472.
24. Weiss O, Jimenez-Montano MA, Herzel H. Information content of protein sequences. *J Theor Biol*. 2000;206(3):379-386.
25. Pande VS, Grosberg AY, Tanaka T. Nonrandomness in protein sequences - evidence for a physically driven stage of evolution. *Proc Natl Acad Sci U S A*. 1994;91(26):12972-12975.
26. Mackenzie CO, Zhou JF, Zheng F, Grigoryan G. A tertiary alphabet for the observable protein structural universe captures sequence-structure relationships. *Protein Sci*. 2016;25:75-76.
27. Szoniec G, Ogorzalek MJ. Entropy of never born protein sequences. *Springerplus*. 2013;2(1):200
28. Peto M, Kloczkowski A, Honavar V, Jernigan RL. Use of machine learning algorithms to classify binary protein sequences as highly-designable or poorly-designable. *BMC Bioinform*. 2008;9(1):487. http://dx.doi.org/10.1186/1471-2105-9-487.
29. Munteanu CR, Gonzalez-Diaz H, Borges F, de Magalhaes AL. Natural/random protein classification models based on star network topological indices. *J Theor Biol*. 2008;254(4):775-783.
30. Kabat EA, Wu TT. The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of -sheets in concanavalin A. *Proc Natl Acad Sci U S A*. 1973;70(5):1473-1477.
31. Xia X, Xie Z. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Mol Biol Evol*. 2002;19(1):58-67.
32. Borguesan B, Inostroza-Ponta M, Dorn M. NIAS-server: neighbors influence of amino acids and secondary structures in proteins. *J Comput Biol*. 2017;24(3):255-265.
33. DasGupta D, Kaushik R, Jayaram B. From Ramachandran maps to tertiary structures of proteins. *J Phys Chem B*. 2015;119(34):11136-11145.
34. Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteinsextended database. *Nucleic Acids Res*. 2019;47(D1):D475-D481.
35. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28 (23):3150-3152.
36. Bateman A, Martin MJ, Orchard S, et al. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506-D515.
37. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16(6):276-277.
38. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*. 2004;32:W500-W502.
39. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res*. 2019;47(W1):W402-W407.
40. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, et al. FoldIndex([c]): a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*. 2005;21(16):3435-3438.
41. Mezei M. Exploiting sparse statistics for a sequence-based prediction of the effect of mutations. *Algorithms*. 2019;12(10):214-220.
42. Tsygvintsev A. Natural vs. random protein sequences: the novel neural network approach based on time series analysis. *Journal of Proteins and Proteomics*. 2020;11(1):11–16. http://dx.doi.org/10.1007/s42485-020-00029-8.
43. Koepnick B, Flatten J, Husain T, et al. De novo protein design by citizen scientists. *Nature*. 2019;570(7761):390.
44. Lin YR, Koga N, Tatsumi-Koga R, et al. Control over overall shape and size in de novo designed proteins. *Proc Natl Acad Sci U S A*. 2015;112 (40):E5478-E5485.
45. Koga N, Tatsumi-Koga R, Liu G, et al. Principles for designing ideal protein structures. *Nature*. 2012;491(7423):222-227.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kaushik R, Zhang KYJ. A protein sequence fitness function for identifying natural and nonnatural proteins. *Proteins*. 2020;88:1271–1284. https://doi.org/10.1002/prot.25900