Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 1402−1410

pubs.acs.org/jpr

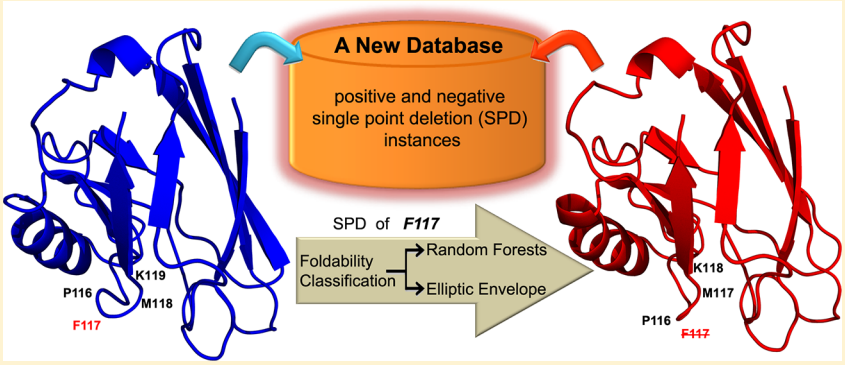
Analyzing Change in Protein Stability Associated with Single Point Deletions in a Newly Defined Protein Structure Database

Anupam Banerjee,[†] Yaakov Levy,[§] and Pralay Mitra*[,‡]

[†]Advanced Technology Development Centre and [‡]Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

[§]Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Supporting Information



ABSTRACT: Protein backbone alternation due to insertion/deletion or mutation operation often results in a change of fundamental biophysical properties of proteins. The proposed work intends to encode the protein stability changes associated with single point deletions (SPDs) of amino acids in proteins. The encoding will help in the primary screening of detrimental backbone modifications before opting for expensive in vitro experimentations. In the absence of any benchmark database documenting SPDs, we curate a data set containing SPDs that lead to both folded conformations and unfolded state. We differentiate these SPD instances with the help of simple structural and physicochemical features and eventually classify the foldability resulting out of SPDs using a Random Forest classifier and an Elliptic Envelope based outlier detector. Adhering to leave one out cross validation, the accuracy of the Random Forest classifier and the Elliptic Envelope is of 99.4% and 98.1%, respectively. The newly defined database and the delineation of SPD instances based on its resulting foldability provide a head start toward finding a solution to the given problem.

KEYWORDS: *protein modification, amino acid InDel, protein foldability prediction, feature analysis*

1. INTRODUCTION

Amino acid insertions/deletions (InDels) and mutations in the protein sequence may alter protein structure. However, as InDels result in the readjustment of the protein backbone, they introduce substantial leaps in the protein fitness landscape and are considered as a critical facilitator of the evolution process.[1,2] Perturbations in the protein structure are often accompanied by noticeable changes in the biophysical properties, substrate specificity, and registry shifts in case of defined secondary structures.[3] InDels can act as a causative agent for many Mendelian disorders,[4,5] cystic fibrosis,[6] leukemia, and other types of cancers.[7] In spite of being a crucial evolutionary modification process and the reason behind multiple human diseases, the study of protein stability and function due to InDels is one of the most challenging and less explored protein engineering problems. Therefore, it is imperative to analyze and encode the effect of such InDel operations.

Limited experimental studies help us to understand the effect of InDels, on the folding, thermodynamic stability, and specific activity of the proteins. Mostly the InDels were performed on the N- and C-terminal regions of the proteins.[8−11] The deletion of entire domains and its role in the modular protein evolution has been discussed by Weiner et al.[12] Lusetti et al.[13] established the presence of extensive bonded and nonbonded interactions between the C-terminal region and the other parts of the protein by C-terminal deletion mutations of different sizes (ranging from 6 to 25) in the Rec A protein of *Escherichia coli*. The critical role of N- and C-terminal contact in protein folding and stability has been investigated for a family 10 Xylanase protein.[14] Therein the deletion of select terminal residues has been shown to result in loss of stability and function under given experimental conditions. The prevalence of compensatory amino acid

Received: January 19, 2019
Published: February 8, 2019

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

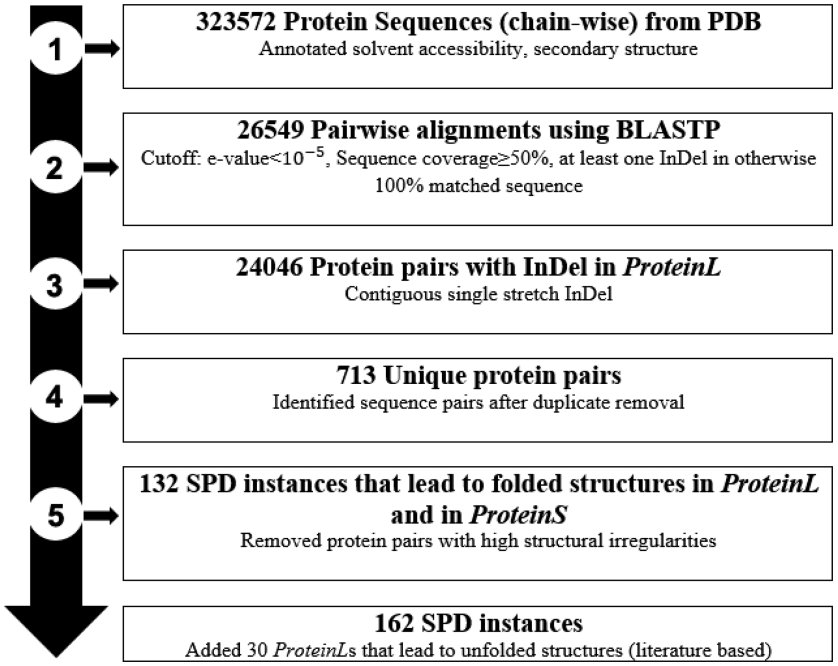


Figure 1. Curation of SPD Database from PDB and literature evidence.

substitutions following InDel operations to arrive at energetically fit conformations have been studied for generic protein structures[2] and Drosophila proteins.[1] In an unusual observation regarding intrinsically disordered proteins, Light et al. suggested that InDel events do not induce disorders but are instead accumulated in intrinsically disordered regions of the corresponding proteins.[15]

Although there are experimental studies that deal with InDels in protein structures, in-silico analysis of the same is limited. Pascarella et al.[16] performed computational analysis on the backbone mutations to deduce that the InDels occurring at loop regions or the N- or C-terminals prefer to be 1−5 residues long. Benner et al.[17] observed that the probability of InDels in a homologous pair of protein sequences was found to increase with an increase in the evolutionary distance between two homologs and thus suggested that a protein of length 30 to 40 amino acids remains, on average, undisrupted by InDels during divergent evolution. An inductive logic programming based machine learning approach was proposed to predict disease-causing nonframeshift (NFS) InDels.[18] The foldability encoding[19] of 72 instances of single point deletions (SPDs) in amino acids of the enhanced green fluorescent protein (eGFP) corroborated well with the experimental findings.[3] The effect of loop length modification on the thermodynamic stability of human muscle acylphosphatase protein was explored using both computational and experimental approaches.[20] In a similar kind of work, the effect of loop length shortening on native state dynamics was also investigated using all-atom molecular dynamics (MD) simulations in the solvent-exposed loop region for four different protein structures.[21] However, a generic computational framework to predict the InDel effect on the thermodynamic stability of any given protein is unavailable until date. The primary impediment is the absence of a standardized database. Unlike Protherm,[22] a mutation database, there are very few experimental findings and no existing database that lists InDels along with their respective changes in thermodynamic stability parameters. The absence of such reportage makes a generic computational framework that predicts the effect of InDels even more elusive.

There are SPD instances described from the helix, sheet, and loop regions that both lead to stable folded conformations and the unfolded state. Lusetti et al. reported that the InDels are most likely to occur in the loop regions of proteins.[13] A majority of loop region InDels may introduce minimum structural changes but some of them result in unfolded conformations, and the observation demonstrates that the loop region InDels are fascinating. In this work, we consider SPDs from all secondary structures with an emphasis on the loop regions. In process, we introduce a new database for the classification and analysis of SPDs in proteins. Our database consists of 162 SPD instances out of which 132 positive instances leading to folded conformations are curated from the Protein Data Bank (PDB)[23] and the remaining 30 SPDs (negative instances) leading to unfolding are identified from existing literature information. For each positive SPD instance, we have protein pair *ProteinL* and *ProteinS*, where *ProteinL* and *ProteinS* are the experimentally derived structures (present in the PDB) without and with the SPD.

We analyze the SPD database to derive crucial insights regarding the structural and physicochemical properties of the SPD instances delineating the folded conformations and unfolded state. Next, we use an adaptive sampling based oversampling technique to mitigate the class imbalance and utilize the distribution of the SPDs in the feature space to prepare simple classification frameworks discerning foldability. We construct a Random Forest (RF)[24] classifier exploiting the distribution of the delineating parameters in the feature space in both the original and the class balanced database. On the class balanced data set, our proposed RF classifier reports an average accuracy of 97% (98% on the original data set) and an average Mathews correlation coefficient (MCC) of 0.90 (0.93 on the original data set) over 100 iterations of three-fold cross-validation (CV). Using the leave one out cross validation (LOOCV) protocol, our RF classifier reports an accuracy of 99.4% with no misclassification of any negative sample when

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

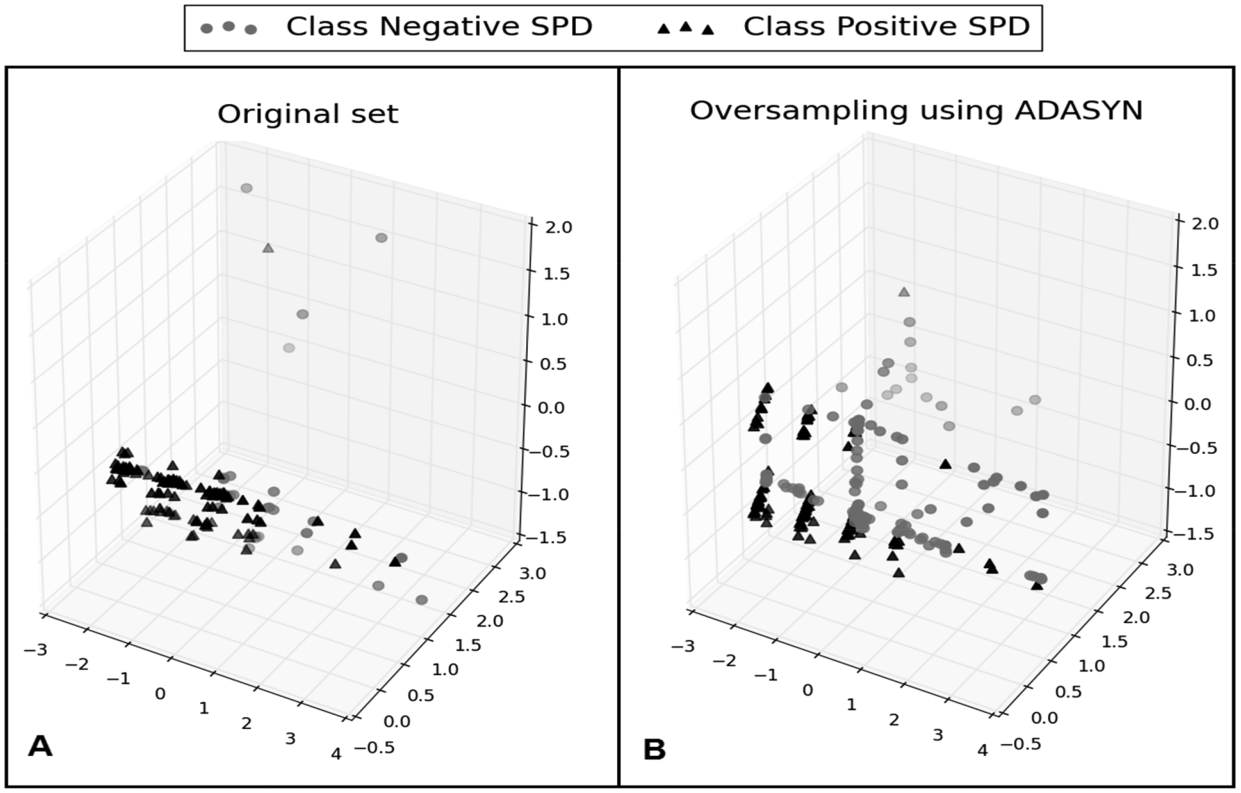


Figure 2. Distribution of positive (black filled triangles) and negative (gray filled circles) data when the feature space is mapped to the three principal components. (A) Classification data (total 162 instances; 132 positives and 30 negatives), (B) augmented classification data (264 instances) with 102 class balancing synthetic data as generated using ADASYN.

trained and tested using both the class balanced and the original data set. Acknowledging the difficulty in obtaining the negative samples, we develop an additional elliptic envelope (EE)[25] based outlier detection system trained on only the positive SPD instances. The accuracy of the system in detecting protein foldability is 98.1% while adhering to the same LOOCV protocol.

The proposed work introduces a new database and two orthogonal foldability classification systems corresponding to SPDs in protein structures. In the absence of any such existing practice, the approaches and algorithms presented provide a head start toward finding a solution to the given problem. The classification framework and the database of SPDs can be downloaded from http://cse.iitkgp.ac.in/~pralay/resources/SPD_Pred/.

2. MATERIALS AND METHODS

2.1. Database Preparation

The SPD instances that result in adequately folded conformations were identified from the PDB using the sequence of steps provided in Figure 1. The curation started by compiling 323 572 single chain protein sequences corresponding to all the PDB entries. We used STRIDE[26] to assign solvent accessibility and secondary structure information for each of the protein chains. Next, we probed for identical protein sequences from our compiled 323 572 chain-wise protein sequence data using BLASTP[27] (V2.2.30). A sequence pair composed of the query sequence and the aligned sequence was considered if the expectation value of BLAST alignment was less than 10^{-5} and sequence coverage was more than 50%.

Barring the insertion/deletion sites, 100% sequence identity between query and the aligned sequence was considered for further shortlisting. Double entries with the interchanged query and aligned sequences were regarded as one for further shortlisting. Out of the 26 549 shortlisted pairs, we identified 24 046 pairs that had a single contiguous deletion in either protein sequence. At this step, multiple deletion cases were omitted, and instances with a single contiguous deletion stretch in either protein sequence were considered. Next, pairwise duplicate entries with the same sequence (or sequence segments) but belonging to different PDB IDs were pruned down. After duplicate removal, we got 713 unique protein pairs where there is deletion in either one of the pairs.

It should be noted that structural irregularities like missing backbone coordinates are an issue for structure-based computations. Thus, proteins with minor structural irregularities were rectified using ITASSER,[28] whereas proteins with significant structural irregularities were removed from our database. Finally, we considered 132 protein pairs with an SPD in either sequence in a pair. As presented in the Introduction, we will refer to each protein pair as *ProteinL* and *ProteinS* where *ProteinL* is the PDB structure without SPD and *ProteinS* is the PDB structure with an SPD in *ProteinL*. These 132 protein pairs are considered as positive samples where a protein can sustain an SPD.

To enable our database for the biclassification problem, we included negative instances, where an SPD in a protein leads to an unfolded state. We identified 30 such cases from the literature. Twenty-eight of them were selected from the analysis of Arpino et al.[3] from the 87 (42 functional and 45 nonfunctional) deletion instances in enhanced green fluo-

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

rescent mutant protein (eGFP; PDB ID: 4EUL). The crystal structure of eGFP consists of a chromophore that forms the protein's characteristic green phenotype. For our present structure-based analysis, we accommodated the original residues: Thr65, Tyr66, Gly67 in place of the chromophore by using the I-TASSER[28] web server. The remaining two instances of SPDs were considered as per experimental evidence.[14] The article discusses the role of N- and C-terminal residues in protein stability and folding of a family 10 Xylanase (BSX) (PDB ID: 2FGL) under extreme conditions and reports that the removal of Trp6 and Tyr343 affects the *in vivo* folding and activity of the protein.

A detailed summary of the 162 SPD instances is included in Supplementary Tables S1 and S2. The SPD database contains proteins from five different classes as classified by the SCOP[29] database. Lengthwise, the shortest protein is 23 residues long, and the longest protein consists of 605 residues. This substantiates the diversity in our database and justifies the robustness of our classification and analysis. The entire database of SPDs composed of the curated PDB files for both *ProteinL* and *ProteinS* along with a description of each SPD is available at http://cse.iitkgp.ac.in/~pralay/resources/SPD_Pred/ for further research. We also maintain a script that will automatically update our SPD database in synchronization with the PDB update.

2.2. Class Balancing Using ADASYN

Most classification algorithms expect balanced data sets, and consequentially, the presence of imbalanced class distributions may fail to represent the distinguishing characteristics of the database adequately. In the present context, we performed oversampling of the negative SPD samples in the defined feature space using the Adaptive Synthetic (ADASYN)[30] sampling approach for imbalanced learning. ADASYN generated more synthetic data corresponding to minority samples that were difficult to learn thereby shifting the classification decision boundary toward them. The use of K nearest neighbor (in our case $K = 7$) to generate synthetic negative samples ensured that the generated samples successfully represents the distribution of the negative data set. Figure 2 depicts the augmented data set and that of the original data set following principal component analysis and mapping the feature space to the three principal components. The imbalanced-learn open-source python toolbox[31] was used for the present implementation. The positive (132) and the negative (30) SPD instances were used as input to the ADASYN framework. The distribution of the SPDs in the feature space was utilized to generate 102 additional synthetic negative SPD instances.

In the following sections, we will refer experimentally verified SPD instances (162 including 132 positive and 30 negative) as the classification database, and the ADASYN enhanced instances (264 with additional 102 negative cases) as augmented classification database.

2.3. SPD Site Features

We considered a total of 11 features to describe the structural, evolutionary and physicochemical environment of the residue involved in the SPD of *ProteinL*. The *weighted contact number* (*WCN*)[32,33] measure reflects the density of the neighboring residues with respect to the residue in consideration and is also an indicator of the contact number of a residue. The *evolutionary conservation score* (*ECS*) represents how well a residue is conserved in its position concerning the structural homologs of *ProteinL*. Considering the residue facing SPD, the *aromatic cluster score* (*ACS*) indicates the size of the aromatic cluster to which the residue belongs. The *hydrophobic core score* (*HCS*) and the *hydrophobic buried core score* (*HBCS*) measure the percentage of residues that are hydrophobic and are solvent inaccessible (along with hydrophobic) out of the total number of residues present in the vicinity of a given residue. The *chemical bond information* (*CBI*) lists the number of hydrogen bonds, ionic bonds, and disulfide bonds formed by the residue facing deletion. The *long-range contact order* (*LRCO*) reflects the number of long-range interactions involving the residue under consideration. Finally, the *hinge residue and flexible residue information* (*HFRI*) as proposed by Emekli et al.[34] has been used to identify whether the residue undergoing SPD is a part of a hinge or a flexible region. The detailed formulations of all the features are provided in the Supporting Information.

2.4. Random Forest Based Classification

The RF classifier uses an ensemble of decision tree classifiers constructed on various subsamples of the database. A diverse set of decision tree classifiers are created by considering random subsets of the feature space. The randomness introduced during the process controls the overfitting and improves predictive accuracy. We considered scikit-learn[35] implementations of the RF algorithm for the classification. The algorithm in its present form considers the average of the probabilistic prediction of each of the classifier to predict the class of a sample. Our RF confidence score measures the proportion of decision trees voting in favor of the desired class label and the proportion of decision trees voting against it.

We considered both stratified three-fold CV and LOOCV approaches to assess the efficacy of the classification algorithm. The cross-validated classification was carried out using only the experimentally verified instances in the augmented classification database.

2.5. Elliptic Envelope for Outlier Detection

The EE[25] based outlier detection technique was adopted to construct an outlier detection framework that would serve as an alternative in the absence of sufficient negative samples. The idea was to establish the SPD instances that point to folding as inliers which in turn would help us to identify the negative SPD instances as outliers. In the present case, the positive SPD instances (the inliers) are considered from diverse protein structures, and the outlier detection system helps in preparing a classifier based on only the distribution of the inliers in the feature space. While the negative SPD instances are only used for testing, the classification of the positive instances concerning the distribution of the inliers also informs us about the fitness of the classifier. The elliptic envelope was fit to the central data points (constituting the inliers) based on a robust covariance estimate computed from the distribution of the SPDs in the feature space. The EE implementation from scikit-learn[35] was used to construct the outlier detector, while the cross-validation of the one-class classifier was carried out sticking to the LOOCV protocol.

3. RESULTS AND DISCUSSIONS

3.1. Inferences from SPD Database

We considered a set of 162 SPD instances (Table S2) to classify the foldability of the protein conformations resulting from backbone modifications. Out of the 162 instances, for 132 SPD instances, we have protein structures with and

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

without the SPD available in the PDB. Whereas, for the remaining 30 instances, literature evidence informs the existence of a protein structure (in PDB), whose SPD leads to unfolding. Interestingly, the majority of the SPDs reported are in the nonterminal regions. Corroborating with the observation of Lusetti et al.,[13] the majority of the positive SPDs is reported from the loop regions (88 cases), whereas 31 and 13 cases are located at the helix and sheet regions, respectively. For negative SPDs, the data distribution is 2, 20, and 8 at the helix, sheet, and loop regions of the protein structures. Overall, 59% SPDs are reported at the protein loop regions.

While analyzing the database, we found that there was no evidence of positive SPD in which glutamine was deleted. In the case of loop regions, none of the SPD data was reported that involved cysteine or glutamine. The distribution of amino acids in the tolerated SPDs (Figure 3A) indicates that mostly

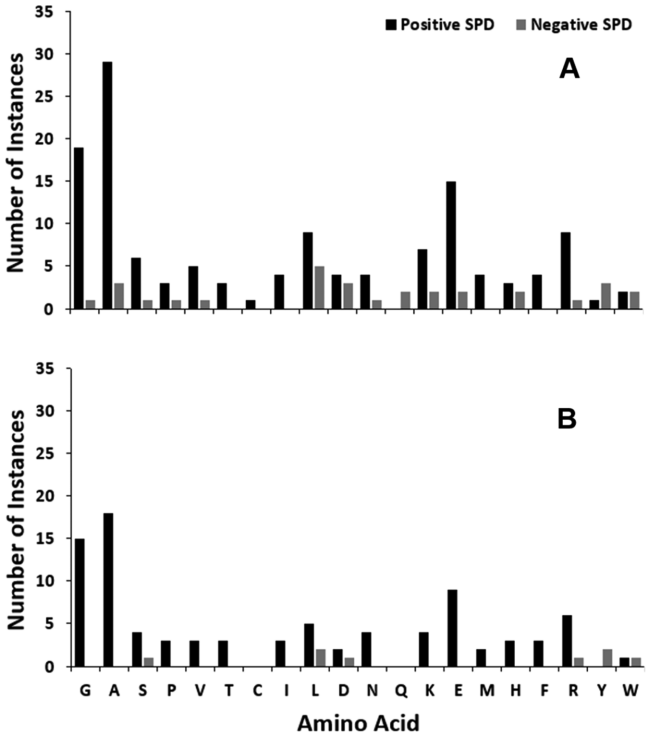


Figure 3. Distribution of amino acids in positive and negative SPD instances (A) for the entire SPD database and (B) for SPD at loop regions.

alanine (22%), glycine (14.4%), and glutamic acid (11.4%) were deleted from a protein. The same trend is reflected for the loop region SPDs, where mostly alanine (20.4%), glycine (17%), and glutamic acid (10.2%) were deleted. The remaining amino acids were each reported in less than 10% cases in the entire database. In the case of negative SPDs, the deletion of leucine (16.7%) was maximally reported. While for the loop region SPDs, in which the deletions were not accepted, we noted two instances each of leucine and tyrosine, and one instance each of tryptophan, aspartic acid, serine, and arginine (Figure 3B). Fascinatingly, for the positive SPDs, we observed that in 39 (out of 132) cases the deleted amino acid was the same as the amino acid preceding or succeeding it.

Considering the solvent accessibility (SA) of the deleted residue, we found that 17 (11), 81 (19), and 34 (0) positive (negative) SPD examples were in the buried (SA ≤ 9%), intermediately buried (9% < SA < 64%), and exposed (SA ≥ 64%) environment, respectively. We identified that chance of foldability is least when the SPD takes place at the buried regions, and all the proteins facing SPDs folded if it is in the exposed area.

We assessed the segregation capability of the *WCN*, *ECS*, *HBCS*, and *LRCO* features using two population hypothesis testing (2-tail tests)[36] before classifying the foldability of proteins subject to SPDs. For all the four features, the null hypothesis H_0 suggests that the population mean μ_1 of the SPD instances that lead to folding is equal to the population mean μ_2 of the instances of SPD that lead to unfolding. The alternative hypothesis H_1 suggests that they are unequal. The summary of the two population hypothesis testing (considering both $\sigma_1^2 = \sigma_2^2$ and $\sigma_1^2 \neq \sigma_2^2$) is provided in Table 1. The

Table 1. Two Population Hypothesis Testing (2-Tail Tests)

	$\sigma_1^2 = \sigma_2^2$, $\alpha = 0.05$			$\sigma_1^2 \neq \sigma_2^2$, $\alpha = 0.05$		
features	DoF[a]	$\lvert t_{\alpha,n_1+n_2-2}\rvert$	$\lvert t_0\rvert$	~DoF[b]	$\lvert t_{\alpha,\nu}\rvert$	$\lvert t_0{}^*\rvert$
WCN	160	1.97	4.81	114	1.98	7.80
ECS	160	1.97	7.58	59	2.00	9.46
HBCS	160	1.97	7.15	38	2.02	6.22
LRCO	160	1.97	8.37	144	1.98	14.79

[a]Degrees of freedom (DoF). [b]Rounded fractional DoF.

hypothesis testing strongly asserted in favor of the alternative hypothesis and confirmed the segregation capability of the selected features. Our hypothesis testing revealed that the SPD of residues with higher *WCN*, higher *ECS*, higher *HBCS*, and lower *LRCO* can result in a possible unfolding of the protein structure. Out of the four features, *LRCO* turns out to be the best indicator of protein foldability considering SPD sites in *ProteinL*. Further, we present the box plots of 4 selected features chosen to construct the binary classifier discerning foldability (Figure 4). The box plots of the same features corresponding to SPDs in the loop region are provided in Figure S1. The *ACS* and *HBI* features have not been considered for hypothesis testing as the measure does not follow a normal distribution. However, as the box plots suggest, SPDs associated with an aromatic cluster or with more numbers of hydrogen bonds have a greater tendency of unfolding. For SPDs in the loop region, considering the *hinge residue* feature, we found that two out of eight negative SPDs was a hinge residue whereas the count was nine out of 88 for positive SPDs. In case of *flexible residue information*, none of the negative SPDs belonged to a flexible loop region, whereas in 24 out of 88 instances the residue fitted to a flexible loop region for positive SPDs. The data justify that proteins with SPDs in a flexible loop region have a greater chance of folding.

3.2. Classifying Protein's Foldability Subject to SPD

Once we have established the segregation capabilities of the chosen features, we classified the deletion instances as folding or nonfolding with the help of simple classification frameworks. To alleviate the problems arising from the lesser number of negative SPD instances, we used the augmented classification database to construct an RF-based classifier and prepared an EE based outlier detector using only the positive SPD instances.

The RF classifier is usually immune to overfitting[37] that makes it quite efficient to work with small databases. Initially,

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

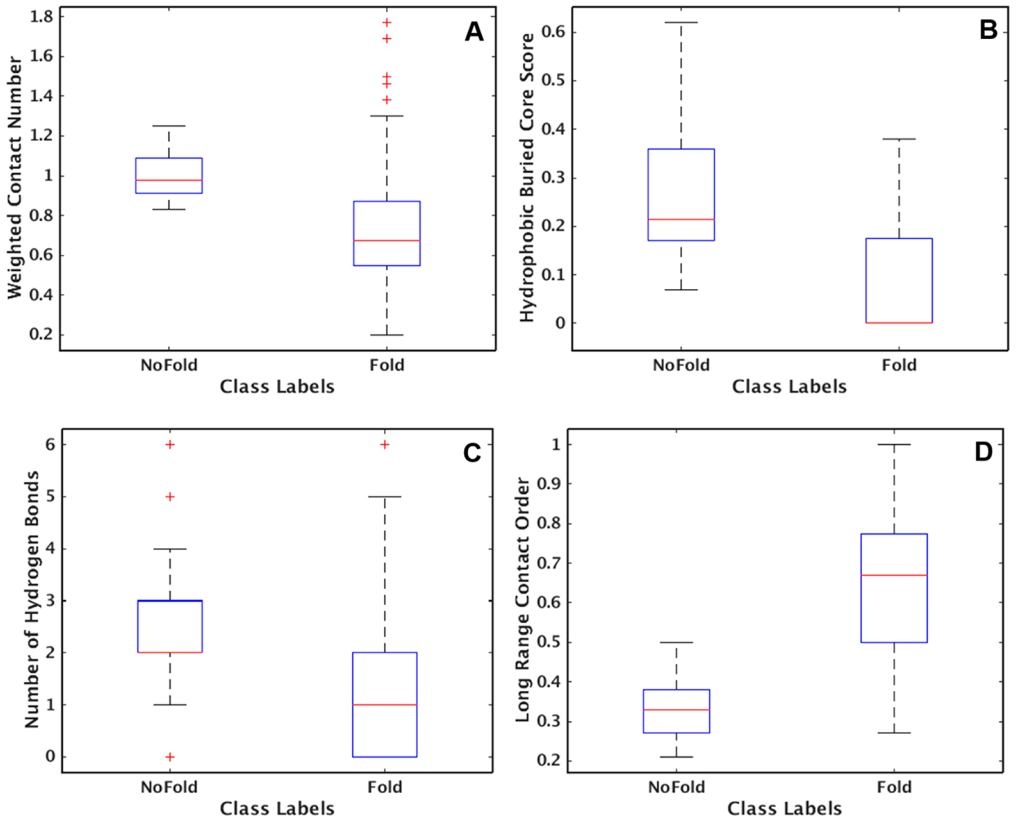


Figure 4. Distribution of different features among (A) SPDs with lower WCN have a higher chance of folding, (B) SPDs with higher HBCS have a higher chance of unfolding, (C) SPDs with higher HBI have a higher chance of unfolding, (D) SPDs with lower LRCO have high chances of unfolding.

Table 2. Evaluation of Classification and Outlier Detection with LOOCV

		positive (%)		negative (%)			
training data set	classification method	true	false	true	false	accuracy (%)	misclassified SPD instances
augmented classification database	random forest	99.2	0	100	0.8	99.4	P114
experimental data set	elliptic envelope	97.7	0	100	2.3	98.1	P73, P82, P126

we performed a stratified three-fold cross-validation (CV) using RF classifier on the augmented classification database. During CV, the database was randomly partitioned into three parts with proportional representation from the positive and negative instances. The third part was tested based on the model developed using the other two parts. We carried out 100 iterations of three-fold CV to minimize the bias resulting out of the random partitioning of the training and test databases. The MCC measure, which varies from −1 to +1, is regarded as a reliable measure even in situations of a higher class imbalanced database. Considering the relatively smaller size of the training set, the result of our classifier on the classification data (MCC score: 0.90 and accuracy: 97%) is encouraging. There were only five instances (P73, P81, P114, P132, and N1) out of 162 that were misclassified in more than 20 iterations of three-fold CV. Training was performed using both the original and oversampled instances, whereas testing was carried out only on the experimentally verified data.

Misclassified instance N1 reports the deletion of a tryptophan (Trp6) in the N-terminal of 10 Xylanase leads to the unfolding of the protein. We observed the existence of an aromatic cluster comprising of Trp6, Phe4, and Tyr343 residues. The aromatic cluster involving long-range interactions provides stability to the protein and is a significant feature that determines the final folded conformation of the protein. Hence, it might be intuitive that removal of bulky Trp6 breaks the aromatic cluster which in turn destroys the protein. *ACS* is an important feature and was included as one of the SPD site features but was not represented well in the smaller training set created due to three-fold CV.

To address the inadequacy arising out of a smaller training set in three-fold CV as well as for better generalization error estimation, we carried out LOOCV on the RF classifier using the augmented classification database. Without misclassification of any negative instance, the RF classifier reported a classification accuracy of 99.4% while adhering to the LOOCV protocol (Table 2). The number of decision trees in the RF voting in favor of correct positive classification was 0.93 and in favor of a negative classification was 0.81 when averaged over the 192 LOOCV classifications. The result is a good indication of the robustness of our features and the generalizability of the RF-based classification framework. We wish to report that while using the RF classifier, the accuracy increased a little if we did not augment our database. For three-fold CV, the increase was 1% due to the lesser number of negative instances in comparison to the augmented database. The averaged MCC

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

score repeated over 100 iterations for three-fold CV increased to 0.93 while using the experimental classification database. Still, the average number of decision trees in the RF voting in favor of a correct positive classification remained the same whereas the average number of decision trees voting in favor of a negative classification reduced to 0.72. The result indicates that the augmented classification database aids in more robust classification.

Considering the difficulty in curating the negative SPD instances, we developed an outlier detection system using only the positive SPDs from the classification database. The outlier detection framework fits the distribution of the features in the positive database in an EE. This framework relies solely on the distribution of the data in the feature space and serves as a robust orthogonal classification system, which in addition to the RF classifier can be used to check the foldability of SPDs in a given protein structure. The result of the outlier detection framework with only three misclassified instances (P73, P82, and P126) indicates good accuracy (Table 2).

3.3. Analyzing Classification

The segregation capability of the individual features has been discussed in the previous section and the combined capacity of the features to delineate the positive SPDs from the negative ones is reflected by the robustness of the RF and EE based classification frameworks. The feature vectors of the synthetic negative samples added by the ADASYN sampling approach in the augmented classification database give rise to a well-balanced database. The RF classifier makes use of this balanced database to prepare an ensemble of decision trees and aggregate the probabilistic prediction of each tree to predict the class of a sample eventually. While assessing the importance of the individual features, we found that the *LRCO* measure contributes maximally (31%) to the delineation process. The deletion of a residue participating in too many long-range interactions makes a protein particularly vulnerable to unfolding. In accordance with the findings of hypothesis testing, the *WCN*, *ECS*, and *HBCS* measure also contributed 20%, 26%, and 9%, respectively, to the delineation process. The remaining features working in tandem with the four measures scaled up the classification accuracy of the RF classifier to 99.4% (while adhering to LOOCV). Distinct distribution of the positive and negative SPD instances was observed while performing hypothesis testing of the *WCN*, *ECS*, *HBCS*, and the *LRCO* measures (see the previous section). These measures along with the remaining features were used to represent the distribution of the positive SPD instances in the EE based outlier detection system which reported an average classification accuracy of 98.1%.

The P13 instance, corresponding to the deletion of the His73 residue of the actin protein, reports a high *WCN* and a low *LRCO* value quite similar to the negative SPD examples. The participation of the histidine in two hydrogen bonds and the presence of glutamate and two aspartates in the residues structural vicinity indicate toward a possible ionic interaction. However, the deletion of His73 does not lead to unfolding. The RF-based classification framework, as well as the EE based outlier detection system; both correctly predicted the foldability of the P13 instance. The foldability of the N1 instance, misclassified in multiple iterations of three-fold CV of the RF classification framework, was anticipated precisely by the RF classification framework adhering to the LOOCV protocol. The augmented classification database leads to an adequate representation of the *ACS* measure hence leading to the correct classification. Interestingly in N1, the EE based outlier detector also correctly predicts the unfolding of 10 Xylanase resulting due to the deletion of Trp6 from the N-terminal region. These two techniques following two entirely distinct approaches together give rise to a robust classification framework.

At the individual level, we found that P114 is being misclassified by the RF classifier. We observed that the *WCN*, *HBCS*, and the *LRCO* scores for P114 (deletion of Ala222 in Green Fluorescent Protein) are similar to the negative SPD instances. However, in P114, alanine is deleted, and as observed earlier, the deletion of alanine is well tolerated in SPD instances. However, the P73, P82, and P126 entries were wrongly misclassified by the outlier detection framework. Considering the consensus of the two frameworks, we find that no SPD instance is being misclassified by both the classifiers. Regarding the classification of SPDs in the loop region, the RF classifier misclassifies none whereas the outlier detection framework misclassifies the foldability of two (P73 and P82) instances.

4. CONCLUSION

A single point InDel or a mutation may cause a pivotal change in the stability and functionality of any protein structures. However, the impact of InDel operations on a protein structure is relatively less explored as opposed to the effect of mutation operations. The absence of databases listing the impact of SPDs on the thermodynamic stability of a protein and lack of computational frameworks to infer the effect of such InDels are significant impediments in primary screening before opting for expensive in vitro experiments. The lack of study makes the problem even more elusive. Nevertheless, the importance of the problem remains pertinent.

We curated the entire PDB to prepare a database consisting of 132 instances of SPDs that lead to folded conformations and listed another 30 SPD instances that point to unfolding. Given the constraints due to the limited information regarding the deletion instances, it allowed us to establish basic classification frameworks distinguishing the two classes of data. The deletion of a residue in the presence of the same amino acid preceding or succeeding it in 39 out of 132 instances and the presence of the deleted residue in the flexible loop region of the protein in 24 out of 88 instances, all resulting in folded conformations indicated toward exciting patterns in the database. The *WCN*, *ECS*, *HBCS*, and *LRCO* all performed well on two-population hypothesis testing and were well delineated in the positive and negative SPD instances.

The database was further used to prepare two frameworks: an RF classifier utilizing both the positive and negative SPD samples and an EE based outlier detection system using only the positive SPD instances. The RF classifier ensured minimum overfitting and was trained on the class balanced database. The RF classifier with LOOCV reports high accuracy (99.4%) without any misclassified negative sample. Even the EE based outlier detection system trained on positive examples shows high accuracy (98.1%) without any misclassification on negative data. The foldability classification of the SPDs in a given protein structure with the help of two orthogonal classifiers indicates a robust classification framework. We believe that our database can be enriched with the new experimental findings, and the approaches enlisted address

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

significant issues associated with such structural modifications of the protein backbone.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00048.

Details on methods and results; SPD site features; distribution of different features among SPDs in loop region; details of 162 *ProteinLs* and 132 *ProteinSs* considered for database; SPD classification database (P1−P132 and N1−N30) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pralay@cse.iitkgp.ac.in. Phone: +91-3222-282344.

ORCID

Anupam Banerjee: 0000-0002-2859-7705
Yaakov Levy: 0000-0002-9929-973X
Pralay Mitra: 0000-0003-4119-3788

Author Contributions

Conceived: Y.L., P.M. Designed the experiment: A.B., P.M. Performed the experiments: A.B. Analyzed data: A.B., P.M. Wrote the paper: A.B., Y.L., P.M.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the University Grants Commission under India-Israel Joint Research Project−2014 (Sanction letter F.NO. 6−6/2014(IC), Dt. 16−10−2014).

■ REFERENCES

(1) Leushkin, E. V.; Bazykin, G. A.; Kondrashov, A. S. In Insertions and deletions trigger adaptive walks in Drosophila proteins. *Proc. R. Soc. B*; The Royal Society, 2012; pp 3075−3082.

(2) Tóth-Petróczy, Á.; Tawfik, D. S. Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol. Biol. Evol.* **2013**, *30* (4), 761−771.

(3) Arpino, J. A.; Reddington, S. C.; Halliwell, L. M.; Rizkallah, P. J.; Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **2014**, *22* (6), 889−898.

(4) Bamshad, M. J.; Ng, S. B.; Bigham, A. W.; Tabor, H. K.; Emond, M. J.; Nickerson, D. A.; Shendure, J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **2011**, *12* (11), 745.

(5) MacArthur, D. G.; Balasubramanian, S.; Frankish, A.; Huang, N.; Morris, J.; Walter, K.; Jostins, L.; Habegger, L.; Pickrell, J. K.; Montgomery, S. B.; et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **2012**, *335* (6070), 823−828.

(6) Collins, F. S.; Drumm, M. L.; Cole, J. L.; Lockwood, W. K.; Woude, G. V.; Iannuzzi, M. C. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **1987**, *235* (4792), 1046−1049.

(7) Ye, K.; Wang, J.; Jayasinghe, R.; Lameijer, E.-W.; McMichael, J. F.; Ning, J.; McLellan, M. D.; Xie, M.; Cao, S.; Yellapantula, V.; et al. Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* **2016**, *22* (1), 97.

(8) Wang, X.-C.; Yang, J.; Huang, W.; He, L.; Yu, J.-T.; Lin, Q.-S.; Li, W.; Zhou, H.-M. Effects of removal of the N-terminal amino acid residues on the activity and conformation of firefly luciferase. *Int. J. Biochem. Cell Biol.* **2002**, *34* (8), 983−991.

(9) Pastor, A.; Singh, A. K.; Shukla, P. K.; Equbal, M. J.; Malik, S. T.; Singh, T. P.; Chaudhuri, T. K. Role of N-terminal region of Escherichia coli maltodextrin glucosidase in folding and function of the protein. *Biochim. Biophys. Acta, Proteins Proteomics* **2016**, *1864* (9), 1138−1151.

(10) Lu, Z.; Wang, Q.; Jiang, S.; Zhang, G.; Ma, Y. Truncation of the unique N-terminal domain improved the thermos-stability and specific activity of alkaline α-amylase Amy703. *Sci. Rep.* **2016**, *6*, 22465.

(11) Ubaid-ullah, S.; Haque, M. A.; Zaidi, S.; Hassan, M. I.; Islam, A.; Batra, J. K.; Singh, T. P.; Ahmad, F. Effect of sequential deletion of extra N-terminal residues on the structure and stability of yeast iso-1-cytochrome-c. *J. Biomol. Struct. Dyn.* **2014**, *32* (12), 2005−2016.

(12) Weiner, J.; Beaussart, F.; Bornberg-Bauer, E. Domain deletions and substitutions in the modular protein evolution. *FEBS J.* **2006**, *273* (9), 2037−2047.

(13) Lusetti, S. L.; Wood, E. A.; Fleming, C. D.; Modica, M. J.; Korth, J.; Abbott, L.; Dwyer, D. W.; Roca, A. I.; Inman, R. B.; Cox, M. M. C-terminal deletions of the Escherichia coli RecA protein. Characterization of in vivo and in vitro effects. *J. Biol. Chem.* **2003**, *278* (18), 16372−80.

(14) Bhardwaj, A.; Leelavathi, S.; Mazumdar-Leighton, S.; Ghosh, A.; Ramakumar, S.; Reddy, V. S. The critical role of N-and C-terminal contact in protein stability and folding of a family 10 xylanase under extreme conditions. *PLoS One* **2010**, *5* (6), e11347.

(15) Light, S.; Sagit, R.; Sachenkova, O.; Ekman, D.; Elofsson, A. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol. Biol. Evol.* **2013**, *30* (12), 2645−2653.

(16) Pascarella, S.; Argos, P. Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* **1992**, *224* (2), 461−471.

(17) Benner, S. A.; Cohen, M. A.; Gonnet, G. H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **1993**, *229* (4), 1065−1082.

(18) Bermejo-Das-Neves, C.; Nguyen, H.-N.; Poch, O.; Thompson, J. D. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). *BMC Bioinf.* **2014**, *15* (1), 111.

(19) Jackson, E. L.; Spielman, S. J.; Wilke, C. O. Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. *PLoS One* **2017**, *12* (4), e0164905.

(20) Dagan, S.; Hagai, T.; Gavrilov, Y.; Kapon, R.; Levy, Y.; Reich, Z. Stabilization of a protein conferred by an increase in folded state entropy. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (26), 10628−10633.

(21) Gavrilov, Y.; Dagan, S.; Levy, Y. Shortening a loop can increase protein native state entropy. *Proteins: Struct., Funct., Genet.* **2015**, *83* (12), 2137−2146.

(22) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **2004**, *32* (suppl_1), D120−D121.

(23) Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.* **2017**, *1607*, 627−641.

(24) Breiman, L. Random forests. *Machine learning* **2001**, *45* (1), 5−32.

(25) Rousseeuw, P. J.; Driessen, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **1999**, *41* (3), 212−223.

(26) Heinig, M.; Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **2004**, *32* (suppl_2), W500−W502.

(27) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410

(28) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2015**, *12* (1), 7−8.
(29) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247* (4), 536−540.
(30) He, H.; Bai, Y.; Garcia, E. A.; Li, S. In ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, Neural Networks. *IEEE World Congress on Computational Intelligence*; IEEE, 2008; pp 1322−1328.
(31) Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Machine Learn. Res.* **2017**, *18* (1), 559−563.
(32) Lin, C. P.; Huang, S. W.; Lai, Y. L.; Yen, S. C.; Shih, C. H.; Lu, C. H.; Huang, C. C.; Hwang, J. K. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Struct., Funct., Genet.* **2008**, *72* (3), 929−935.
(33) Marcos, M. L.; Echave, J. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* **2015**, *3*, e911.
(34) Emekli, U.; Schneidman-Duhovny, D.; Wolfson, H. J.; Nussinov, R.; Haliloglu, T. HingeProt: automated prediction of hinges in protein structures. *Proteins: Struct., Funct., Genet.* **2008**, *70* (4), 1219−1227.
(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Machine Learn. Res.* **2011**, *12* (Oct), 2825−2830.
(36) Hines, W. W.; Montgomery, D. C.; Borror, D. M. G. C. M. *Probability and Statistics in Engineering*; John Wiley & Sons, 2008.
(37) Gislason, P. O.; Benediktsson, J. A.; Sveinsson, J. R. Random forests for land cover classification. *Pattern Recognition Letters* **2006**, *27* (4), 294−300.

DOI: 10.1021/acs.jproteome.9b00048
J. Proteome Res. 2019, 18, 1402−1410