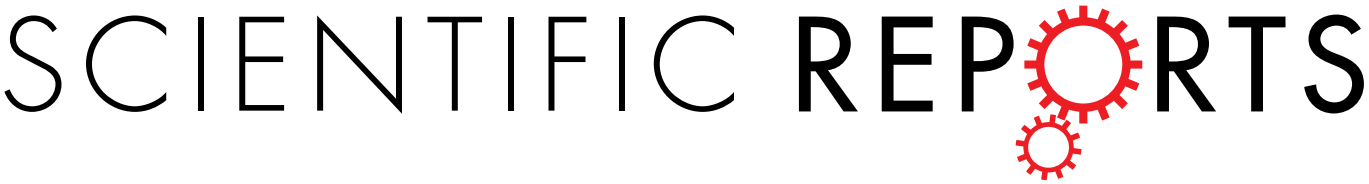


OPEN

Complete fold annotation of the human proteome using a novel structural feature space

Received: 04 January 2017
Accepted: 14 March 2017
Published: 13 April 2017

Sarah A. Middleton[1], Joseph Illuminati[2] & Junhyong Kim[1,3]

Recognition of protein structural fold is the starting point for many structure prediction tools and protein function inference. Fold prediction is computationally demanding and recognizing novel folds is difficult such that the majority of proteins have not been annotated for fold classification. Here we describe a new machine learning approach using a novel feature space that can be used for accurate recognition of all 1,221 currently known folds and inference of unknown novel folds. We show that our method achieves better than 94% accuracy even when many folds have only one training example. We demonstrate the utility of this method by predicting the folds of 34,330 human protein domains and showing that these predictions can yield useful insights into potential biological function, such as prediction of RNA-binding ability. Our method can be applied to *de novo* fold prediction of entire proteomes and identify candidate novel fold families.

Although protein sequences can theoretically form a vast range of structures, the number of distinct three-dimensional topologies ("folds") actually observed in nature appears to be both finite and relatively small[1]: 1,221 folds are currently recognized in the SCOPe (Structural Classification of Proteins—extended) database[2], and the rate of new fold discoveries has diminished greatly over the past two decades. Nevertheless, extending the catalog of protein fold diversity is still an important problem and fold classifying the entire proteome of an organism can lead to important insights about protein function[3–5]. Large-scale fold prediction typically involves computational methods, and the computational difficulty of *ab initio* structure prediction has led to template matching (e.g., using methods such as HHPred[6]) as the most common method for predicting the structure. When sequence-based matching is difficult, other fold recognition approaches must be employed, such as protein threading. Threading-based methods, especially those that combine information from multiple templates, have been among the most successful algorithms in recent competitions for fold prediction[7,8], but are bottlenecked by long run times. Machine learning-based methods have also been used, which can be designed either to recognize pairs of proteins with the same fold[9,10] or classify a protein into a fold[11,12]. Although these methods have shown promising results for a subset of folds, they have so far not been able to generalize to the full-scale fold recognition problem. This failure can mainly be attributed to the severe lack of training data available for most SCOPe folds, as well as the highly multi-class nature of the full problem, which requires distinguishing between over 1,000 different folds[12].

Here we introduce a method for full-scale fold recognition that integrates aspects of both threading and machine learning. At the core of our method is a novel feature space constructed by threading protein sequences against a relatively small set of structure templates. These templates act as "landmarks" against which other protein sequences can be compared to infer their location within structure space. We show the utility of this feature space in conjunction with both support vector machine (SVM) and first-nearest neighbor (1NN) classifiers, and further develop our 1NN classifier into a full-scale fold recognition pipeline that can predict all currently known folds. Applied to the entire human proteome, our method achieves 95.6% accuracy on domains with a known fold and makes thousands of additional high-confidence fold predictions for domains of unknown fold. We demonstrate utility by inferring new functional information, focusing on RNA-binding ability. The structure and function annotations of the entire human proteome are provided as a resource for the community.

[1]Genomics and Computational Biology Program, University of Pennsylvania, Philadelphia, PA 19104, USA. [2]Department of Computer Science, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. Correspondence and requests for materials should be addressed to J.K. (email: junhyong@sas.upenn.edu)

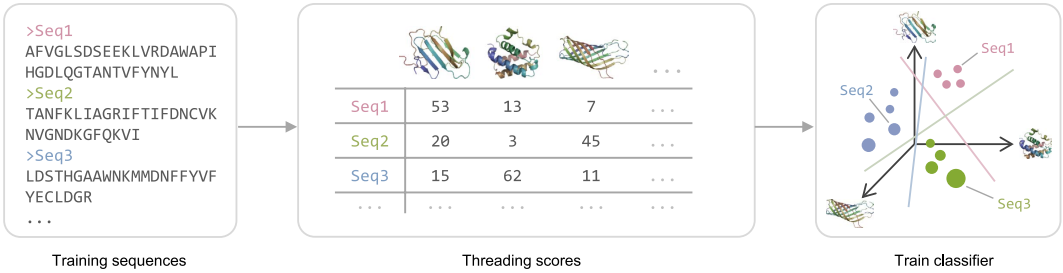


Figure 1. Overview of PESS construction. Training sequences of known fold are threaded against a set of structure templates, and the resulting threading scores act as coordinates within a structural feature space (the PESS). A classifier can then be trained to recognize the subspace occupied by each fold in the PESS. Different colors indicate the fold of each sequence and are shown here only for visualization.

Results

The protein empirical structure space (PESS). Our approach is based on the idea of an empirical kernel[13], where the distance between two objects is computed by comparing each object to a set of empirical examples or models. We have previously applied this idea to RNA secondary structure analysis[14], and we show here that it can be adapted to proteins. The objects being compared are amino-acid sequences and the distance we would like to compute is similarity of tertiary structure. We selected a set of 1,814 empirical threading templates that describe the three-dimensional coordinates of atoms of proteins of known structures. We use only a small subset of known structures for our template library which we find sufficient to construct an informative structural distance function. Using the threading templates we mapped amino-acid sequences to a structural feature space, where the coordinates of each sequence reflect its threading scores against the templates (see Methods). We refer to this as the protein empirical structure space (PESS). Using the PESS, we trained a classifier to recognize every fold (Fig. 1). Since protein domains are the unit of classification in SCOPe, we applied this approach to protein domains as units rather than full proteins.

Fold recognition performance. We tested the PESS in combination with 1NN or SVM classifiers (Fig. 2a and b) using three popular benchmarks from the TAXFOLD paper[12]. These benchmarks are designed to test the ability of a method to distinguish between increasing numbers of folds: 27 folds in EDD, 95 in F95, and 194 in F194. Each fold has at least 11 training examples. The accuracy of our classifiers are shown in Table 1 along with the results reported by several other published methods[12,15–19]. Since some of the sequences in these benchmarks are similar to the templates that make up our feature space (that is, similar to the sequences of the domains from which the templates are derived), we ran our classifiers both with and without filtering of these template-similar sequences (Table 1; "filtered" versions correspond to benchmarks where sequences with >25% pairwise identity with any template were removed; see Methods). Performance was almost identical with and without filtering, indicating that similarity of training and/or testing sequences with the feature space templates did not have a major effect on classification. Our SVM classifier performed the best on all three benchmarks, with the exception of the EDD dataset, where the best performance was from the method of Zakeri *et al.* when it was used in combination with known Interpro functional annotations. Our 1NN classifier also performed very well on all three benchmarks, outperforming all but our SVM on F95 and F194. We note that some of these publications used slightly modified versions of the benchmarks, which may affect the comparison (see Methods for details). We next asked whether our method actually performed better than simply using the top-scoring template from the 1,814 that make up our feature space. We found that directly using the fold of the top template as the fold prediction resulted in 52.1, 56.4, and 57.4% accuracy on EDD, F95, and F194 respectively. Therefore, using the threading scores as a feature space rather than for direct classification improved performance considerably.

The benchmarks described above included only a subset of the 1,221 folds in SCOPe v.2.06. Recognizing all folds simultaneously is challenging; not only is it a highly multiclass problem, but it also suffers from a lack of training examples for a large fraction of the folds. We focused on our 1NN classifier, which requires only a single training example per fold, to scale to the full fold recognition task. To train the classifier to recognize a larger number of folds, we downloaded domain sequences from SCOPe v.2.06 that had been pre-filtered to ≤25% pairwise identity and split this into training and testing sets (see Methods, "SCOP datasets and final classifier"). We refer to this as the SCOP-25 test. The same 1,814 templates were used to extract features as before, and any training or testing sequences that had >25% identity with one of these templates were removed from the set. The training set consisted of 5,686 sequences covering 760 folds in classes "a" through "g", and the test set consisted of 1,171 sequences covering 271 folds. A 1NN classifier trained on this training set achieved 94.1% accuracy on the test set (precision = 0.917, recall = 0.882). Using a combined SVM+1NN classifier (see Methods) did not improve performance (acc = 94.0%, precision = 0.857, recall = 0.880), indicating that the 1NN classifier alone is sufficient for good classification on this dataset. Overall, the SCOP-25 test shows that even when there is very low sequence identity between training and test examples (≤25%) and many possible folds for a test sequence to be classified into (760), performance is still good. Notably, of the 760 folds in the training data, 82% (622) had fewer than 10 training examples, and 43% (325) were "orphan" folds with only one training example. Accurate classification into these folds is expected to be particularly difficult due to the small amount of training data. To determine how well our method performs relative to the number of training examples, we calculated precision and recall separately

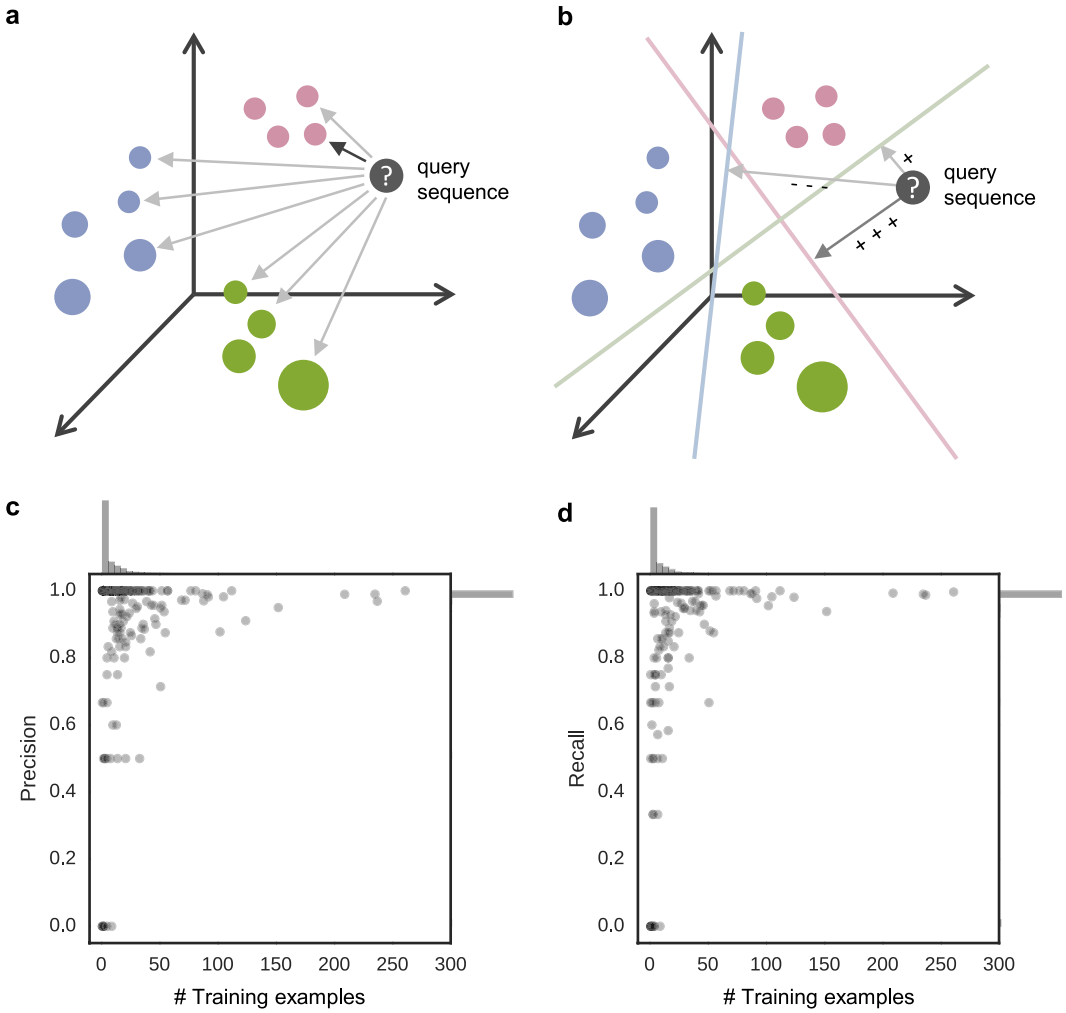


Figure 2. Classification and performance using the PESS. (**a** and **b**) Two different methods of classification using the PESS. Colored circles represent training examples within the PESS and are colored by fold. (**a**) In 1NN classification, the PESS distance between the query (gray circle) and all training examples is computed and the query is assigned to the fold of the nearest training example (dark gray arrow). (**b**) In 1-vs-all SVM classification, the PESS distance between the query and each of the fold-level hyperplanes (dotted lines) is computed, and the query is assigned to the fold that gives the best score (dark gray arrow), based on signed distance from the fold's hyperplane. (**c**) Precision and (**d**) recall measures were computed for each fold separately after 1NN classification of the SCOP-25 set using the PESS and plotted against the number of training examples for each fold. Marginal histograms show the distribution of folds along each axis.

for each fold (Fig. 2b and c). Focusing specifically on orphan folds, for which classification should be most difficult, we found that the classification accuracy for these folds was 79.0% (precision = 0.724, recall = 0.766). Thus, our method can recognize folds reasonably well even when there is only a single training example.

Due to the stringent filters imposed on the training and test set described above, this test still only covered a little over half of the 1,221 folds in SCOPe. In order to cover more folds, we created a second training and test set derived from SCOPe domains pre-filtered to ≤40% pairwise identity, which is the cutoff used in several other papers when testing fold recognition[10,12,20–22]. We refer to this as the SCOP-40 test. We note that in order to get representation of all 1,221 folds in the training set, we did not remove examples that were similar to template sequences (see minimal effects noted for TAXFOLD above). We split the domains into a training and test set (see Methods, "SCOP datasets and final classifier"), resulting in a training set of 7,413 sequences in 1,220 folds (one fold, g.2, could not be covered because its domain sequences were too short), and a test set of 6,321 sequences in 609 folds. Using a 1NN classifier trained on this training data (again, using the same 1,814 templates to extract features), the test sequences were classified with 97.3% accuracy (precision = 0.960, recall = 0.944). The classification on orphan folds in this set was 95.6% (precision = 0.932, recall = 0.939). The better overall performance on the SCOP-40 test than the SCOP-25 test can likely be mainly attributed to the more relaxed pairwise identity threshold in the SCOP-40, but nonetheless, these results show that the 1NN classifier can scale up to the full set of SCOPe folds and maintain high accuracy for folds with only a single training example.

Method		EDD	F95	F194
Dehzangi *et al.*[a]		88.2	—	—
Saini *et al.*[a]		86.6	—	—
Lyons *et al.*[a]		93.8	—	—
Zakeri *et al.*		88.8/96.9[b]	—	—
Yang and Chen		90.0	82.4	79.6
Wei *et al.*[c]		92.6	83.6	78.2
This method (PESS)	1NN – filtered[d]	89.9	84.6	82.6
	1NN – all[e]	90.6	84.6	82.5
	SVM – filtered[d]	95.9	92.3	90.7
	SVM – all[e]	95.7	91.9	90.5

Table 1. Overall % accuracy on three benchmarks using 10-fold cross validation. [a]Using a slightly modified EDD set with 21 additional domains (3418 total) (see Methods). [b]With Interpro functional annotations. [c]Using modified versions of EDD (3625 domains), F95 (6791 domains), and F194 (8525) (see Methods). [d]Using a filtered version of the benchmarks which removed any examples with >25% pairwise identity with a template (based on the sequence of the domain on which the template was based). [e]Using the full benchmark sets. Some training or testing sequence may be similar or identical to templates.

Proteome-scale fold prediction of human proteins. The ability of the PESS to accurate recognize all folds with relatively little threading makes it well suited for classifying large, proteome-scale datasets. Here we applied our new method to predicting the fold of protein domains curated from the entire human proteome. Since the 1NN-only classifier performed better than the SVM+1NN combined classifier on the full-scale fold recognition test, we used the 1NN-only classifier (trained on 1,220 SCOPe folds, as above) to predict the folds of all human protein domains.

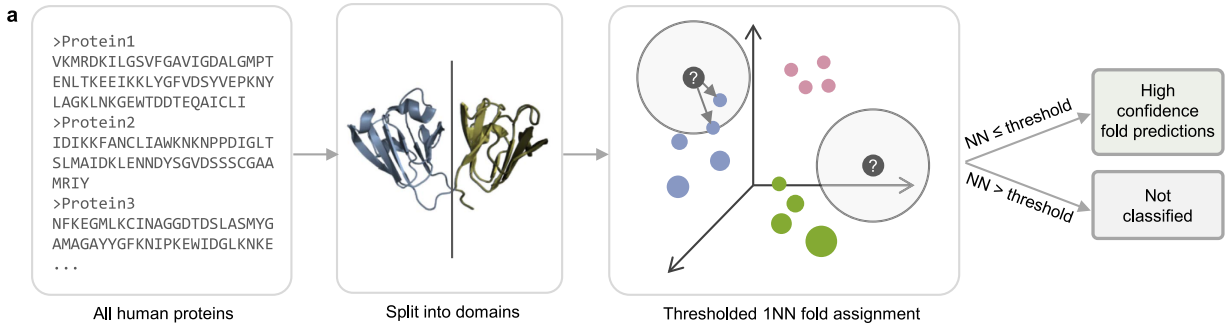
An overview of our whole proteome fold classification pipeline is shown in Fig. 3a. In contrast to SCOP-derived benchmarks, whole proteomes present several additional challenges for fold recognition. One of the major bottlenecks is the process of segmenting whole proteins into domains, which is often slow and error-prone. We did not attempt to address this issue here, but instead make use of the existing domain segmentation of the human proteome performed by the Proteome Folding Project[5]. Another challenge is recognizing domains that do not belong in any of the known fold categories, e.g. due to segmentation errors, being disordered, or belonging to a previously undiscovered fold. To address this problem, we defined a distance threshold for classification based on the typical distance between a domain and its nearest neighbor when the true fold of the domain is not represented in the feature space (see Methods). When a query domain's nearest neighbor is farther than this threshold distance, the domain is assigned to a "no classification" category (Fig. 3a).

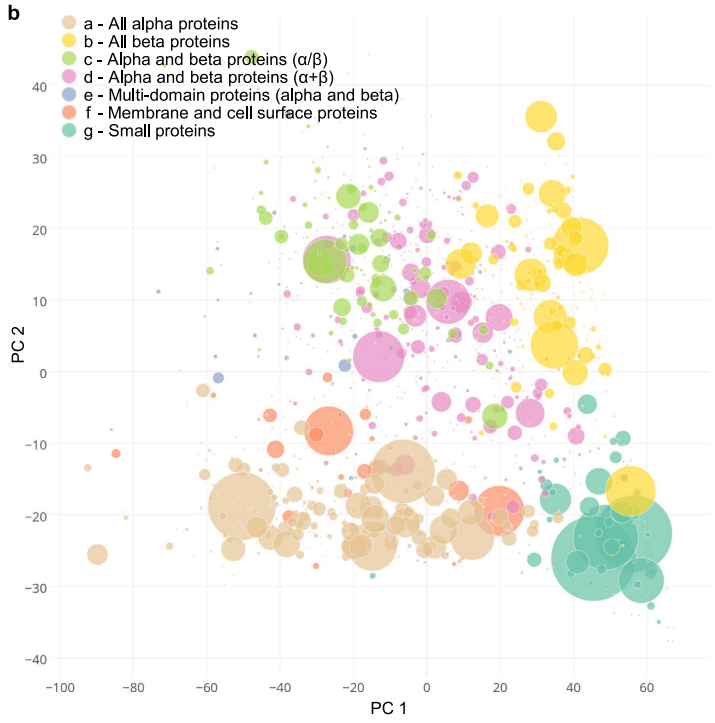
There were a total of 34,330 human domains with length greater than 30 residues in the Proteome Folding Project dataset, corresponding to 15,619 proteins. Of these, 20,340 domains (59%) had a nearest neighbor within the distance threshold and were classified into an existing fold by our method. Only 128 of these domains were previously placed into a fold with high confidence by the Proteome Folding Project[5]. To test how well our predictions match with what is currently known about human protein structures, we used a blastp search against PDB to identify 2,211 human domain sequences with a "known" fold; that is, an identical or highly similar PDB entry with a SCOPe fold classification. Our classifier made a fold prediction for 1,873 (84.7%) of these domains, and 95.6% of these predictions exactly matched the known SCOPe fold.

Overall, 757 of the 1,221 SCOPe folds had at least one human domain predicted by our method. The distribution of domains across folds was highly skewed, with the majority of folds having only a few predicted domains and a small number of folds having many (Fig. 3b). This agrees with previous observations that domains are not evenly distributed in protein structure space[1,23]. The top 10 folds accounted for 38.9% (7,908) of the classified domains, and the most common fold (Beta-beta-alpha zinc fingers) alone encompassed 9.1% (1,853) of the fold predictions (Fig. 3c). A full list of fold predictions is provided in Supplementary Table S1.

Human RNA-binding proteins. RNA-binding proteins (RBPs) are an important class of proteins that function in almost all aspects of RNA biology, including splicing, translation, localization, and degradation. It would be valuable to fully define which folds have potential RNA binding function and use this information to improve our annotations of RBPs. We obtained a list of 1,541 currently known RBPs in humans from a recent RBP census[24] and extracted the corresponding domains from our dataset. There were 1,816 domains with fold predictions, matching 243 different folds.

Since not every domain in an RBP is expected to actually bind RNA, we first sorted these folds into "likely RNA-binding domain (likely RBD)" and "likely auxiliary" groups. The RBPs in the census were primarily identified based on hits to a list of Pfam families with RNA-binding function, so we defined the likely RBD folds as those with at least two RBP domains with a hit ($E < 0.01$) to this RNA-binding Pfam list. There were 720 such domains which encompassed 78 different folds. The most common folds included several with well characterized RNA-binding function, such as Ferredoxin-like, which includes the RNA recognition motif (RRM); Eukaryotic type KH-domain (KH-domain type I); and dsRBD-like (Fig. 3d). Next, we defined the auxiliary folds as those with at least one RBP domain but fewer than two hits to the RNA-binding Pfam list. By this criteria, we identified





c

Fold	Description	# Human Domains
g.37	Beta-beta-alpha zinc fingers	1853
a.118	Alpha-alpha superhelix	960
a.4	DNA/RNA-binding 3-helical bundle	844
g.39	Glucocorticoid receptor-like	762
d.92	Zincin-like	633
f.23	Single transmembrane helix	623
b.34	SH3-like barrel	595
b.72	WW domain-like	573
a.140	LEM/SAP HeH motif	535
g.5	Midkine	530

d

Fold	Fold Description	# RBDs	# Dom Total
d.58	Ferredoxin-like	212	321
g.37	beta-beta-alpha zinc fingers	43	1853
a.118	alpha-alpha superhelix	43	960
b.40	OB-fold	31	164
d.51	Eukaryotic type KH-domain	28	32
a.4	DNA/RNA-binding 3-helical bundle	23	844
b.38	Sm-like fold	21	24
d.50	dsRBD-like	17	24
b.34	SH3-like barrel	17	595
g.66	CCCH zinc finger	16	23

Figure 3. Fold classification of the human proteome. (**a**) Overview of classification process. Full length human protein sequences were split at predicted domain boundaries to create one or more separate domain sequences per protein (Drew *et al.* [5]). Domain sequences were mapped to the PESS and classified by 1NN classification. A threshold was applied to the nearest neighbor distance (dotted circle), whereby only domains with a nearest neighbor closer than the threshold distance were classified. (**b**) PCA projection of fold centroids within the PESS, scaled by number of human domains predicted to belong to that fold. Centroids were calculated based on the location of each fold's training examples within the PESS and are colored by SCOP class. (**c**) Top ten folds by number of human domain predictions. (**d**) Top ten likely RNA-binding folds, ranked by number of confirmed RNA-binding domains (RBDs). Confirmed RBDs were determined based on matches to a curated list of RNA-binding related Pfam families.

165 folds, the most common being the Cytochrome C fold (14 domains) and RING/U-box E3 ligase fold (12 domains). These folds are likely to represent other functions performed by the RBPs; however, we note that the lack of a Pfam match does not preclude RNA-binding function, so some of these auxiliary folds may in fact be RNA-binding.

The RBP census contained 21 cases where a protein was known to bind RNA but the type of RBD was not yet identified. Using our method, we matched three of these RBPs to one or more of the likely-RBD folds established above. One of these RBPs was Fam120a (also called C9orf10), which was previously found to have RNA-binding activity at its C-terminal end, but the type of RNA binding domain was not determined[25]. Our method predicted a DNA/RNA-binding 3-helical bundle fold within the RNA-binding region of this protein. Loosening the classification threshold slightly (NN distance ≤ 20) allowed us to identify potential RBDs for three more of the RBPs, including a partial Ferredoxin-like fold at the N-terminal of Int8 and a PABP domain-like fold in Int10.

We next looked to see if there were any additional proteins represented in the likely-RBD folds that were not already annotated as being RBPs by the census. We found 6,249 such proteins, which overlapped substantially

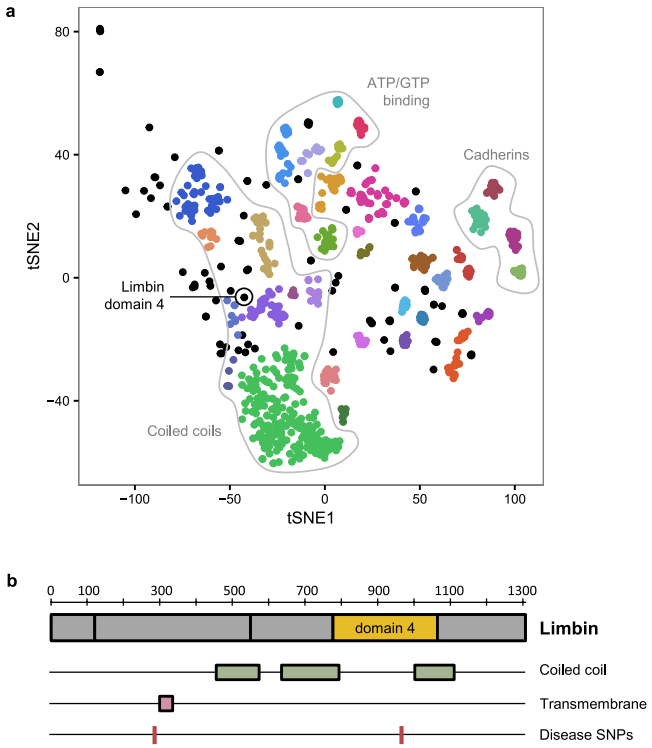


Figure 4. Analysis of unclassified human domains. (**a**) t-SNE projection of human domains with nearest-neighbor distance ≥30. Colors indicate cluster assignment by DBSCAN; unclustered domains are shown in black. Dotted lines show related groups of domains. (**b**) Overview of the *EVC2* protein product, Limbin, and its known structure elements. The location of the domain with a putative novel fold is shown in yellow.

with a recently published set of 6,657 novel RBP predictions by RBPPred (1,981 overlapping genes not previously annotated as RBPs)[26]. The ~2,000 concordant predictions by these two orthogonal methods more than double the number of previous RBP annotations[24]. We note that for many of our RBP predictions, we cannot confidently predict their RBP status based on fold alone because some of the likely-RBD folds have other functions besides RNA-binding (e.g. some superfamilies of the Ferredoxin-like fold can be protein binding instead of RNA binding), which may explain some of the non-overlapping predictions between our method and RBPPred. Nonetheless, several of the likely-RBD folds appear to be highly enriched in known RNA-binding domains, suggesting that functional annotation transfer is possible for these folds. For example, of the 32 domains predicted by our method to have the KH-domain fold, only four did not have a hit to the RNA-binding Pfam list, and of these, three were already known to be KH-domain RBPs based on the RBP census. The one domain that was not in the census was part of the Blom7 protein (also called KIAA0907), which has an experimentally determined structure (PDB: 2YQR) that confirms structural similarity to the KH-domain, despite the lack of a Pfam match. A full list of our new RBP predictions and likely-RBD folds is in Supplementary Table S2.

Novel folds in the human proteome. Each year at least a few new folds are added to SCOPe (e.g. 13 new folds were added in the latest release). As noted above, there were ~14,000 human protein domains, or ~40% of domains, that were not assigned to known folds. While some of these might be due to problems of segmentation, we hypothesize many of them represent uncharacterized folds. As a preliminary analysis of potential novel folds in the human proteome, we extracted a set of human domains that were not close to any of our training examples (NN distance ≥ 30) and clustered them (Methods). This resulted in 36 clusters (Fig. 4a and Supplementary Table S3), which we examined for evidence of novel folds.

We first looked for incorrect domain boundary prediction or errors of our prediction method. Many of the domains were unusually long (>500 residues) compared to the average domain in the training set (195 residues), suggesting that they may in fact be multiple domains. For example, there were four neighboring clusters that contained almost exclusively domains from the Cadherin family of proteins. Most of these domains were longer than 500 residues and overlapped multiple repeats of the Cadherin motif based on Pfam annotations. The Cadherin fold is modeled as a single repeat in SCOPe, so this is likely a case where fold classification failed due to improper domain definition. A similar problem was observed for six clusters containing domains from several different classes of ATP/GTP binding proteins, where each domain spanned multiple distinct Pfam annotations that are likely to represent separate folds. Overall, we found that 26 of the clusters were potentially the result of such segmentation errors.

The largest cluster contained 208 domains, most of which were of a reasonable length (289 residues on average). On closer examination, we found that a large fraction of these domains were predicted to have a coiled coil

structure. The SCOPe hierarchy places most coiled coil domains in a separate class (class H) that was not included in the training data. Therefore, this cluster can possibly be explained by the absence of the correct fold within our training data, although it is not truly novel. Eight other neighboring clusters were also found to have predominantly coiled coil structure, indicating that these structures can potentially explain a substantial fraction of our unclassified domains.

We also examined the un-clustered domains, which might be isolated examples of novel folds. One domain, the fourth predicted domain of the protein Limbin (residues 775–1067), was found not to overlap any known Pfam, SCOP, or other structural annotation. Although this domain was located in the feature space in proximity to the coiled coil clusters (Fig. 4a), it is predicted to be only partially coiled coil (Fig. 4b). We performed a more thorough template search for this domain using HHPred[27], RaptorX[28], and SPARKS-X[29] webservers, but did not identify a significant template match. Limbin is the protein product of the gene *EVC2*, which is involved in the hedgehog signaling pathway and is frequently mutated in Ellis-van Creveld syndrome[30,31]. Interestingly, one of the mutations linked to this disease is found within our domain of interest (Arg870Trp; rs137852928)[30], suggesting that this region is functionally important. Whether this region represents a truly new fold will require additional analysis, but overall these results support the idea that the PESS can be used to identify novel structure groups.

Discussion

Here we have demonstrated the utility of an empirically derived structural feature space composed of threading scores (the PESS) for addressing the problem of fold recognition. The most important characteristics of such a multi-dimensional feature space are the ability to combine characteristics of multiple fold templates for fold recognition and the ability to potentially identify entirely novel folds through interpolation of the feature space. Many types of classifiers can be used in conjunction with this feature space; we showed here that linear SVM achieved good performance on benchmarks where at least 10 training examples were available per fold, and 1NN worked well in the more general case to recognize all known folds. We applied our method to the human proteome, predicted high confidence fold classifications for 20,340 domains, and showed that these predictions can be used to make functional inferences as illustrated by the class of RNA-binding proteins. A distinct advantage of the PESS is that it only requires a single training example per fold when used in conjunction with a 1NN classifier, allowing us to make predictions for all currently known folds in SCOPe. This is critical, since almost half of all SCOPe folds have only one training example in SCOP-20. Another advantage of the 1NN classifier is that adding new training data does not require re-training the whole classifier, making it simple to update the model as new data become available.

One of the limitations of methods that rely on threading is the large amount of time the threading process takes. Threading against all PDB templates can take hours or even days per domain, depending on the computational resources available. In our method, we save time by only threading against representative templates. Nonetheless, threading is still the major time bottleneck, with a single average-sized (200 residue) domain taking 26 ± 2.5 minutes to thread against the 1,814 templates on one CPU core. To make this more feasible for genome-sized datasets, which typically have thousands or tens of thousands of domains, we have implemented an option for parallel processing of the input sequences. Another possible way to decrease the threading time would be to reduce the number of templates in our library. Preliminary results indicate that, depending on the classifier used, the feature space can be substantially reduced with only a minor impact on classification accuracy. In fact, given our framework, we hypothesize that we can create feature spaces at different scales such that threading can be applied in a hierarchical sequence.

The relationship between the structure of macromolecules to their function is a key annotation principle for computational inference. As the number of solved examples increase, we hypothesize that data-driven feature extraction coupled with machine learning methods as in our method and also in methods like deep learning[32], will have high utility in extending whole genome/proteome annotations.

Methods

Feature extraction and classification. Features were created for each input sequence by threading the sequence against a library of 1,814 structure templates to produce a vector of 1,814 threading scores. These scores represent the compatibility of the sequence with each template structure. Each score is directly used as a numerical coordinate within the feature space, which we call the Protein Empirical Structure Space (PESS). Threading was done using CNFalign_lite from the RaptorX package v.1.62[22,33]. This program outputs a raw threading score for each query-template pair that is calculated from the optimal alignment of the query sequence and the template[22,33]. The template library was the default library provided by RaptorX. These 1,814 templates represent a wide range of different structures with low redundancy, but do not necessarily represent all known folds.

Training sequences were threaded against the templates and the resulting scores were normalized by z-standardization. Test sequences were threaded and normalized using the normalization parameters derived from the training sequences.

We constructed fold predictors over the PESS using both a first Nearest Neighbor (1NN) classifier and Support Vector Machine (SVM) classifier. For the 1NN classifier, pairwise Euclidean distances between each training and testing sequence were calculated, and each test sequence was classified into a fold by finding the closest training neighbor and transferring its fold label to the test sequence. For the support vector machine (SVM) classifier, a linear SVM was trained using the one-vs-all multiclass approach with the C parameter (which controls the penalization of misclassification during training) set to 1/N, where N is the number of positive examples in a given fold.

We also constructed a joint SVM+1NN classifier to assist in identification of fold classes with very small number of training examples. First, a linear SVM was trained as described above to recognize only folds that had at least 10 training examples ("large folds"). The remaining sequences in the training set ("small folds") were combined into a single class labeled "other", and this class was not used for classification. A separate 1NN classifier

was trained on only the small fold training examples. Classification was then done in two phases: first, all test examples were provided to the SVM, and any test example that received a positive confidence score (based on the signed distance from the hyperplane) was classified into whichever fold gave the highest confidence score; second, the examples that were not classified in the first step were passed to the 1NN model for classification.

All classifiers were implemented in Python using the scikit-learn package[34].

Performance assessment. Prediction accuracy was calculated as the fraction of test examples that were classified into the correct fold. Precision (the number of true positives divided by the sum of the true and false positives) and recall (the number of true positive divided by the sum of the true positives and false negatives) were calculated separately for each fold and averaged across the folds. For both precision and recall, we excluded folds where the denominator was zero for the SCOP benchmark (611 folds excluded for recall calculation; 618 folds excluded for precision calculation).

Benchmark comparison to other methods. We obtained three benchmark datasets (EDD, F94, and F195) from the TAXFOLD paper[12]. Each benchmark contains only domain sequences longer than 30 residues with less than 40% pairwise identity, but each contains a different number of folds: EDD contains 3397 sequences in 27 folds, F95 contains 6364 sequences in 95 folds, and F194 contains 8026 sequences in 194 folds. To create the "filtered" version of each of these benchmarks (where redundancy between the benchmark sequences and our 1,814 feature templates is removed), we first obtained the original sequences used to generate the templates, which is included in the template file. We then used blastp[35] to query each of the benchmark sequences against the database of template sequences and removed any benchmark sequence that had more than 25% identity over at least 90% of their length with one of the template sequences. The number of sequences removed from each benchmark was: 420 for EDD, 832 for F95, and 1,104 for F194. Performance on each benchmark was assessed using 10-fold cross validation, with SVM and 1NN classifiers trained and assessed as described above. We compared our results to the percent accuracies reported in recent publications that used these benchmarks with 10-fold cross validation. Some of these publications used modified versions of the benchmarks. Dehzangi *et al.*, Saini *et al.*, and Lyons *et al.* all used a version of EDD that had the same 27 folds, but 21 extra domains[15,16,18]. This is only a small fraction of the total number of domains in this dataset, so we do not expect this to have a major impact on the results. A more major modification was made by Wei *et al.*, who used the same folds for EDD, F95, and F194, but updated the datasets to have 228, 427, and 499 extra domains, respectively[19]. Based on these numbers of added sequences, we estimate that the maximum performance of Wei *et al.* on the original TAXFOLD datasets would be no more than 98.8%, 89.2%, and 83.1%, respectively. However, since their new dataset still used the same cutoff for pairwise similarity as the original (<40%), it is more likely that their results would be roughly the same for both datasets. Thus the results in Table 1 should be comparable.

SCOP datasets and final classifier. We downloaded domains from the SCOPe database v2.06 pre-filtered to less than 25% pairwise identity (SCOP-25 set; filtering performed by the Astral[2] database). We split these sequences into training and testing sets as follows. First, all sequences with >25% identity with any of the 1,814 templates were removed (using blastp as described in the previous section). Then, to ensure that the training sequences represented the range of structures, we used SCOPe pre-filtered to 20% identity (filtering performed by the Astral database) as a guide to select sequences from the SCOP-25 set to be training sequences. Specifically, any sequence from the SCOP-25 set that was also in the SCOP-20 set was placed in the training set. The remaining sequences were placed in the test set. Finally, a few sequences had to be removed due to being too short (sequences <25 residues caused an error during threading with CNFalign_lite), resulting in a final dataset of 5,686 training and 1,171 test sequences where no two sequences have a sequence identity greater than 25%. Orphan folds were defined as folds with only one training sequence, and accuracy for this subset of folds was calculated based on the classification of test examples that belonged to these folds (26 orphan folds had at least one test sequence) from the full classification task (i.e. where all training and testing sequences were present).

The SCOP-40 test was created in a similar manner as above. All SCOPe domains were downloaded pre-filtered to less than 40% pairwise identity (filtering performed by the Astral database), and was split into training and testing sequences. In this case, redundancy with the templates was not removed. Any sequences in SCOP-40 that overlapped with SCOP-20 were put in the training set and the remaining sequences were put in the test set, except 19 sequences that did not have a corresponding fold representative in the training set. Finally, any sequences < 25 residues were removed, giving a final set of 7,413 training and 6,321 test sequences where no two sequences have a sequence identity greater than 40%. Orphan folds were profiled in the same manner as the SCOP-25 set. In this case, 158 orphan folds had at least one test sequence.

The final classifier (used for all novel analysis) was trained using all SCOP-20 sequences (except those <25 residues), which was 7,632 sequences total. We note that this training set is almost identical to the one used in the SCOP-40 set, except it contained slightly more domain sequences, so we expect overall performance to be highly similar to that set.

Human protein analysis. Protein domain sequences for 94 species from the Proteome Folding Project[5] were downloaded from the Yeast Resource Center public data repository (http://www.yeastrc.org/pdr/pages/download.jsp). To obtain only human sequences, we filtered for protein identifiers marked as "NCBI NR" and had "[Homo sapiens]" in the description. There were a total of 34,330 human domains with length greater than 30 residues, corresponding to 15,619 human proteins.

We classified the domains using the SCOP-20-trained 1NN model. We used an additional nearest-neighbor distance threshold to filter out domains that likely do not belong in any of the represented folds. We determined the threshold nearest-neighbor distance for classification as follows: for each test sequence in SCOP-40, we

calculated the nearest neighbor distance before and after removing all SCOP-20 training sequences that belonged to the same fold as the test sequence. We found that a distance threshold of 17.5 provided a good balance between false positives and false negatives (FPR = 9.27%, FNR = 9.49%). After classification with 1NN, only the domains with a nearest-neighbor distance below this threshold we considered confident fold predictions.

Human domain sequences were mapped to PDB entries using a blastp search of PDB requiring that at least 75% of the sequence length had at least 90% identity with a PDB sequence to consider it a match. PDB matches were then mapped to SCOPe classifications using the dir.cla.scope.txt (v.2.06) annotation file downloaded from the SCOPe website.

RNA-binding proteins. A list of 1,541 known human RBPs was obtained from a recent review[24]. Gene names of the RBPs were matched up to the human protein GIs using the UniProt ID mapping tool, and 1,093 of the RBPs were matched to one or more domains (3,263 domains total). This review also defined a list of 799 Pfam domains with functions related to RNA binding, which we used to filter the 3,263 RBP domains down to those that were most likely to be RNA-binding. Domains were assigned PfamA annotations using hmmscan (http://hmmer.org/). Both a "full-sequence" $E \leq 0.01$ and a "best 1" $E \leq 0.1$ was required for assignment. We compared our novel RBP predictions with the novel predictions from the RBPPred paper[26] on the gene level by mapping UniProt IDs to gene names for each list using the ID conversion tool on the UniProt website. Not all UniProt IDs could be mapped to a gene name. The final unique gene lists contained 6,589 genes for RBPPred and 5,668 genes for our method, which we used to compute the overlap.

Novel folds. We extracted all human domains with a nearest neighbor distance ≥ 30 and performed t-SNE on the PESS projections of these domains using scikit-learn with parameters "perplexity = 10, init = 'pca', random_state = 123". Domains were then clustered using DBSCAN from scikit-learn with parameters "eps = 5, min_samples = 5". Domains and clusters were manually examined for potential boundary prediction errors or previous structural annotations.

Data and Code Availability. Benchmark datasets, training data, and all human fold and RBP predictions are available at http://kim.bio.upenn.edu/software/pess.shtml. The fold classification source code is freely available at the same website or at https://github.com/sarahmid/PESS.

References

1. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420,** 218–223 (2002).
2. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42,** D304–D309 (2014).
3. Kim, S. H. *et al.* Structural genomics of minimal organisms and protein fold space. *J. Struct. Funct. Genomics* **6,** 63–70 (2005).
4. Malmström, L. *et al.* Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol.* **5,** 758–768 (2007).
5. Drew, K. *et al.* The Proteome Folding Project: Proteome-scale prediction of structure and function. *Genome Res.* **21,** 1981–1994 (2011).
6. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins Struct. Funct. Bioinforma.* **77,** 128–132 (2009).
7. Huang, Y. J., Mao, B., Aramini, J. M. & Montelione, G. T. Assessment of template-based protein structure predictions in CASP10. *Proteins Struct. Funct. Bioinforma.* **82,** 43–56 (2014).
8. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5,** 725–38 (2010).
9. Cheng, J. & Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **22,** 1456–1463 (2006).
10. Jo, T., Hou, J., Eickholt, J. & Cheng, J. Improving Protein Fold Recognition by Deep Learning Networks. *Sci. Rep.* **5,** 17573 (2015).
11. Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. & Kim, S. H. Recognition of a protein fold in the context of the SCOP classification. *Proteins Struct. Funct. Genet.* **35,** 401–407 (1999).
12. Yang, J.-Y. & Chen, X. Improving taxonomy-based protein fold recognition by using global and local features. *Proteins* **79,** 2053–64 (2011).
13. Scholkopf, B. & Mika, S. Input space versus feature space in kernel-based methods. *IEEE Trans. Neural Netw.* **10,** 1000–1017 (1999).
14. Middleton, S. A. & Kim, J. NoFold: RNA structure clustering without folding or alignment. *RNA* **20,** 1671–1683 (2014).
15. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A. & Sattar, A. A segmentation-based method to extract structural and evolutionary features for protein fold recognition. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **11,** 510–519 (2014).
16. Saini, H. *et al.* Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. *J. Theor. Biol.* **380,** 291–298 (2015).
17. Zakeri, P., Jeuris, B., Vandebril, R. & Moreau, Y. Protein fold recognition using geometric kernel data fusion. *Bioinformatics* **30,** 1850–1857 (2014).
18. Lyons, J. *et al.* Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles from Hidden Markov Models. *IEEE Trans. Nanobioscience* **14,** 761–772 (2015).
19. Wei, L., Liao, M., Gao, X. & Zou, Q. Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique. *IEEE Trans. Nanobioscience* **14,** 649–659 (2015).
20. Lindahl, E. & Elofsson, A. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295,** 613–25 (2000).
21. Ding, C. H. Q. & Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17,** 349–358 (2001).
22. Ma, J., Wang, S., Zhao, F. & Xu, J. Protein threading using context-specific alignment potential. *Bioinformatics* **29,** i257–65 (2013).
23. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372,** 631–634 (1994).
24. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15,** 829–845 (2014).
25. Tanaka, M. *et al.* A novel RNA-binding protein, Ossa/C9orf10, regulates activity of Src kinases to protect cells from oxidative stress-induced apoptosis. *Mol. Cell. Biol.* **29,** 402–413 (2009).
26. Zhang, X. & Liu, S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics* **btw730** (2017).
27. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).
28. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7,** 1511–1522 (2012).

29. Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27,** 2076–2082 (2011).
30. Galdzicka, M. *et al.* A new gene, EVC2, is mutated in Ellis–van Creveld syndrome. *Mol. Genet. Metab.* **77,** 291–295 (2002).
31. D'Asdia, M. C. *et al.* Novel and recurrent EVC and EVC2 mutations in Ellis-van Creveld syndrome and Weyers acrofacial dyostosis. *Eur. J. Med. Genet.* **56,** 80–87 (2013).
32. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
33. Ma, J., Peng, J., Wang, S. & Xu, J. A conditional neural fields model for protein threading. *Bioinformatics* **28,** i59–i66 (2012).
34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
35. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

Acknowledgements

This paper was funded in part by DOE CSGF (DE-FG02-97ER25308) to S.A.M. and Health Research Formula Funds from the Pennsylvania Commonwealth to J.K., which disclaims responsibility for any analyses, interpretations or conclusions. We would like to thank Kanishka Rao for contributions to an early version of the feature space.

Author Contributions

S.A.M. and J.K. conceived the study and wrote the manuscript. S.A.M. implemented the feature space and classifier, applied it to the human proteome, and interpreted results. J.I. contributed to classifier development and validation. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Middleton, S. A. *et al.* Complete fold annotation of the human proteome using a novel structural feature space. *Sci. Rep.* **7**, 46321; doi: 10.1038/srep46321 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.