

# Supporting Information

## Table of Content

### Supplementary Texts

- Text S1:** The normalized number of effective sequences (*N<sub>eff</sub>*) in MSA
- Text S2:** The comparison of Pfam families in Pfam database and supplemented by metagenome data
- Text S3:** Case studies verified the applicability and interpretability of the targeted MetaSource model
- Text S4:** The construction of “PhylaSource” for guiding the 3D structure prediction supplemented by metagenome
- Text S5.** The construction of “EvauleSource” for predicting the E-values when collecting homology sequences

### Supporting Figures

- Figure S1.** TM-scores of the C-I-TASSER models from 168 proteins benchmark dataset for MSAs with different *N<sub>eff</sub>* values using a base of 2.
- Figure S2.** Accuracy estimation of predicted models using C-score.
- Figure S3.** The comparison of Pfam MSA and Metagenome MSA.
- Figure S4.** C-I-TASSER models for 12 cases from 168 proteins benchmark dataset that has large *N<sub>eff</sub>* but low TM-score.
- Figure S5.** The species richness statistic for four biomes (Fermentor, Gut, Lake and Soil).
- Figure S6.** The top 20 importance features (genus) for the multiple-classified Random Forest model.
- Figure S7.** DeepMSA pipeline for multiple sequence alignment generation.
- Figure S8.** Modeling results of C-I-TASSER utilizing genome and metagenome databases.
- Figure S9.** Head-to-head comparison of the protein folding methods using MetaSource selected biome MSA and combined biome MSA.
- Figure S10.** The construction of the PhylaSource and EvaluateSource based on 964 Pfam families with unsolved structure.
- Figure S11.** For the Gut biome, the statistical result based on country distribution.
- Figure S12.** Data collection flow from Pfam database for training and testing MetaSource and benchmarking the C-I-TASSER.
- Figure S13:** A schematic of contact potential.
- Figure S14.** Workflow for targeted MetaSource model construction.

### Supporting Tables

- Table S1.** Wilcox test results for differentiating each pair of two biomes based on species distribution.
- Table S2.** Summary of C-I-TASSER modeling results for 28 Pfam families which has solved experimental structure.
- Table S3.** The contact precision on the 12 cases in the benchmark dataset that has large *N<sub>eff</sub>* > 16 but with low TM-score shown in Figure S1.

**Table S4.** The statistical result for GO annotations (level 3) which were only detected in single biome for the 964 Pfam families.

**Table S5.** Ten case studies for illustration of the Pfam-biome associations.

**Table S6.** The validation result of the MetaSource for the 204 Pfam families with solved structures.

**Table S7.** Reasons of the C-I-TASSER generated un-foldable model for 29 cases of 204 Pfam validation dataset.

## References

## Supplementary Texts

### Text S1. The normalized number of effective sequences (*N<sub>eff</sub>*) in MSA

The depth of a multiple sequence alignment (MSA) is measured by the normalized number of effective sequence (*N<sub>eff</sub>*) in this work:

$$N_{eff} = \frac{1}{\sqrt{L}} \sum_{i=1}^N \frac{1}{1 + \sum_{j=1, j \neq i}^N I[S_{j,i} \geq 0.8]} \quad (S1)$$

where  $L$  is the length of protein,  $N$  is the number of sequences in the MSA,  $S_{j,i}$  is the sequence identity between the  $j$ -th and  $i$ -th sequences.  $I[S_{j,i} \geq 0.8]$  equals to 1 if  $S_{j,i} \geq 0.8$ , or zero otherwise. Therefore, *N<sub>eff</sub>* is essentially equal to the number of non-redundant sequences (sequence identity < 0.8) in the MSA normalized by the protein length.

### Text S2. The comparison of Pfam families in Pfam database and supplemented by metagenome data

To examine the advantage of using microbiome sequences, we compared the MSAs from the Pfam database and the MSAs built by DeepMSA on metagenome databases on 2,214 Hard Pfam families. To make a fair comparison, we did not directly use the existing profile data in the Pfam database. Instead, we have reconstructed the MSA based on the Pfam family sequences using the DeepMSA program that is that same as what we used in this work. For doing so, we first run DeepMSA for each query sequence, and then used the Hidden Markov Model (HMM) generated at the second step of DeepMSA to search against the Pfam family sequences downloaded from the Pfam database to construct the MSA for the 2,214 Pfam family.

In **Figure S3**, we presented a quantitative comparison of the two sets of MSAs. First, due to the enlarged sequence database (3,643,924 from Metagenome database vs. 1,015,317 from 2,214 Pfam families), the average number of sequences for the Metagenome MSA (1645.85±842.45) is 3.6-fold higher than that of the Pfam MSA (458.58±275.62) (**Figure S3A**). Accordingly, the number of effective sequences (*N<sub>eff</sub>*) of Metagenome MSA (75±36.22) is nearly 3-fold larger than that of Pfam MSAs (25.50±13.25) (**Figure S3C**). Although the average sequence identity to the query for the Metagenome MSAs (46.70±28.65) is higher than that of Pfam MSAs, the former has a higher diversity score (7.62±3.15) than the latter (3.89±1.86) (**Figure S3D**), when measured by the *M<sub>eff</sub>* score used in HHblits (1).

It is natural that searching through a larger sequence database usually costs more CPU time for constructing the MSAs. For example, for the Pfam MSAs, the search space was 0.74 TB and the average search time is 1.42±0.85 hours. For the Metagenome MSA, the search space was 2.4 TB and the average search time is 6.38±2.68 hours (if used without MetaSource). This was one of the reasons that motivated us to develop MetaSource to guide metagenome selections. According to our benchmark validation on the 204 Pfam families with solved structure, the MetaSource could reduce the search time by 3.3 times (=5.44h/1.65h). Thus, although the overall time cost of MetaSource is still slightly higher than the Pfam MSA, the MSA quality and contact accuracy are significantly improved when combining MetaSource with microbiome databases.

### Text S3. Case studies verified the applicability and interpretability of the targeted MetaSource model

Through the case studies, our targeted metaSource model shows a strong applicability and good biological interpretability. Among 964 Pfam families (*N<sub>eff</sub>* over 16 and C-score over -0.25), 10 Pfam families are selected for case studies (**Table S4**).

These Pfam families are selected based on the literature review and the comparison of prediction result (measured by the *N<sub>eff</sub>* score) for four commonly used datasets (Uniref100 (2), IMG(2), Tara Oceans (3) and Metaclust (3)). These four datasets are commonly used to assist the structure and function prediction for unsolved proteins.

First, assisted by Soil biome, PF05120 could be supplemented with sufficient homologous sequences (*N<sub>eff</sub>* score=487.5 and C-score=-0.18). Based on our targeted metasource model, this Pfam family is successfully classified into the Soil biome (accuracy: 0.968). However, the other four commonly used datasets supply insufficient homologous sequences, reflected by the lower *N<sub>eff</sub>* score than the Soil biome used in our research: 32.0, 336.8, 69.0 and 178.9 for Uniref100, IMG+ Uniref100, Tara Oceans+Uniref100 and Metaclust+Uniref100, respectively. This result indicates that our targeted

prediction model can accurately predict that Soil biome could be used to supplement the homologous sequence of PF05120. Furthermore, this prediction result could be interpreted by its unique biological role in Soil biome for PF05120: According to the records in Pfam families, the members in PF05120 are annotated as gas vesicles proteins. These gas vesicles proteins are permeable to ambient gases by diffusion and provide buoyancy, enabling cells to move based on the air-soil interface (4, 5). This protein plays an important role in the communications between different soil microbiome communities(4).

Second, the accuracy and interpretability of our targeted metaSource model could also be proved by other biomes: Among the 964 Pfam families with C-score  $>-0.25$ , PF12652 is successfully classified into the biome of Fermentor by our metasource model (accuracy: 0.975). Actually, measured by a high *Neff* score (305.6) and C-score (-0.16), our protein structure prediction results also confirm this result. However, insufficient homologous could be provided by four datasets for PF12652 (*Neff* score 232.6, 264, 295.2, 299.6 for Uniref100, IMG+ Uniref100, Tara Oceans+ Uniref100 and Metaclust+ Uniref100, respectively). The fermentor-related function of proteins in PF12652 could explain this result: based on the records in PF12652, this Pfam family is related to spore development. The bacteria that enrich the spore development function are closely related to anaerobic fermentation, the main function of fermenters (6).

Finally, based on an investigation of the correctly classified Pfam families, great application prospects have sprung up using our targeted MetaSource model: PF13822 (classified into Soil biome, accuracy: 0.982) is identified as Acyl-CoA carboxylase epsilon subunit, which is involved in the biosynthesis of long-chain fatty acids. The long-chain fatty acids are important for *Rhizobium leguminosarum* Growth and Stress Adaptation (7). PF09828 and PF05425 are two important antibiotics. These two antibiotics shows the resistance to chromate and copper, which are harmful to the agricultural plants and human (8, 9). PF09650 (classified into Soil biome, accuracy: 0.965), is identified as putative polyhydroxyalkanoic acid (PHA) system protein, and could produce the bioplastic (10).

#### **Text S4. The construction of “PhylaSource” for guiding the 3D structure prediction supplemented by metagenome**

It might be interesting to look at the phylum label instead of biome label to train a “PhylaSource” model for guiding the search of homologous sequences from the genome sequences from specific Phyla. To do this, we used the same set of 964 Pfam families as MetaSource used to train the PhylaSource model. Since the metagenomic data does not contain the phylum label, instead of using the biome data from metagenome database, we downloaded all the available Prokaryotic and viral genomes (refseq database, <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>) as the taxonomical database for training the PhylaSource. Here we downloaded Prokaryotic and viral genomes, since we found that over 80% the supplemented sequences from the previous 964 MSAs built from metagenome database can be assigned as those genomes by blast (version 2.7.1) with a strict threshold (E-value  $1E-5$ , sequence identity 90%). The data downloaded from NCBI covers 48 phyla and counts for 736GB data with 718,314 protein sequences. These sequences were divided in to 48 sub-blocks, where each block only contains the sequences belong to one phylum. A “PhylaSource” model was constructed using a multi-class logistic regression model (the python package, sklearn) to predict the relative probability of every phylum for a given Pfam family, where the phylum sample with the highest probability was used for guiding the homologous sequence search. For validation, we selected top-10 predicted phylum databases since a single phylum database is too small (average size=15.3GB) to give sufficient supplement sequences. We tested the number of predicted phylum databases from 1 to 48 phyla (ranked by the relative abundance) and found that the highest accuracy of PhylaSource (80.2%) was achieved when the top-10 phyla were used (Figure S10A).

To further examine the practical usefulness of the PhylaSource model for 3D structure modeling, we predicted the phylum probability distribution and selected the top-10 phyla by PhylaSource to supplement their homology sequence search at the step-3 of DeepMSA. For the 204 test families with solved structures, PhylaSource was able to predict the phyla which resulted in a higher contact accuracy in 69.5% of cases or a higher TM-score in 61.4% of cases, compared to that using all genome sequences. The permutation P-value is 0.001, indicating that the difference is statistically significant.

**Figure S10B** displays the average contact accuracy and TM-score of the C-I-TASSER models when using MSAs collected from the all the genome data from NCBI (named as Phyla data) and the dataset selected by PhylaSource on the 204 test families. It was shown that, although the volume of the sequence database by PhylaSource is much smaller (228 GB/per target and 736 GB/per target

for PhylaSource and phyla data respectively), using the targeted dataset from PhylaSource resulted in a higher contact accuracy (0.488 vs. 0.476) and TM-score (0.617 vs 0.615), which corresponds to a P-value=1.5E-5 and 2.3E-5, respectively, in Student's t-test. These results indicate that the MSA from PhylaSource could help depress the sequences from the "wrong" source Phyla. However, with a limited phylum data, the PhylaSource had a lower accuracy of target phyla prediction and a smaller magnitude of contact/TM-score improvement than the MetaSource, although they both demonstrated a similar level of search space and time reduction of sequence databases.

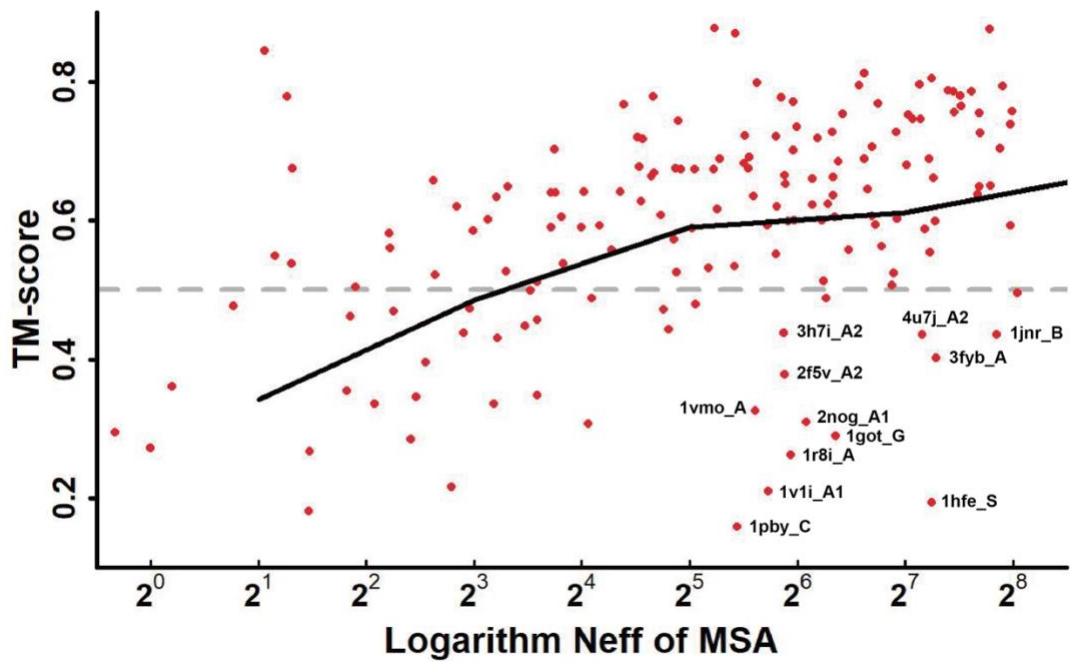
### **Text S5. The construction of "EvaluateSource" for predicting the E-values when collecting homology sequences**

The careful E-value selection in finding homologous sequences is often an important procedure to MSA construction and subsequent 3D structure prediction. Hence, it would be useful to predict an optimal E-value cutoff for collecting the homologous sequences from the metagenome for specific Pfam family, from which the reliable 3D structure would be modeled.

Similar to MetaSource, the EvaluateSource was trained on the 964 Pfam families, where the features for the training set were based on the species distribution for Pfam families, obtained from the Pfam database. We particularly designed a EvaluateSource model to predict the E-value cutoff combination used by hmmer and HHblits in DeepMSA step 3 (for metagenome searching). In the default DeepMSA pipeline (**Figure S7**), the same E-value (=1E-3) was used for the HHblits and hmmer when collecting the homologous sequences from metagenome. In the EvaluateSource pipeline, eight E-values for HHblits (1E+1, 1E+0, 1E-1, 1E-2, 1E-3, 1E-6, 1E-10, and 1E-30) and six E-values for hmmer (1E+1, 1E+0, 1E-1, 1E-2, 1E-3, and 1E-4) were selected as predicted labels. Hence, one of the paired E-values from  $8 \times 6 = 48$  combinations is the final label to be predicted by EvaluateSource when given a Pfam family, and hence 48 MSAs should be constructed for each Pfam families to collect the homologous sequences from metagenome. For each MSA, the sum of the Top  $L$  long-range contact scores are used to estimate the best combination of E-values for HHblits and hmmer for structure prediction, where the E-value cutoff combination associated with the largest contact score would be set as the target label for the training set (11). Finally, four-fold cross-validation shows that the highest accuracy of this model is 82.28% (**Figure S10C**).

To further examine the applicability of this model to 3D structure modeling, we used the same 204 Pfam families with solved structure as the validation dataset. In **Figure S10D**, we compared the modeling results from EvaluateSource with that using default E-value combinations (1E-3 and 1E-3, named as default combination). It was shown that, using the predicted combinations of E-values from EvaluateSource resulted in a slightly higher TM-score (0.613) and contact accuracy (0.508) than that using the default E-values (0.609 and 0.496), which corresponds to a P-value=0.055 for TM-score and a P-value=0.062 for contact accuracy in Student's t-test. These results indicate that the EvaluateSource could help select target E-values for homologous sequence collections, which have resulted in marginal TM-score and contact accuracy improvement. However, EvaluateSouce does not generate similar effect as MetaSource for improving both speed and accuracy of MSA collection and 3D structure prediction. This is probably due to the fluctuation of sequence distances among different protein families, while the inherent linkage between protein families and the ecological species groups could not be captured by the generic sequence distances such as E-value cutoffs.

## Supporting Figures



**Figure S1.** TM-scores of the C-I-TASSER models from 168 proteins benchmark dataset for MSAs with different *Neff* values using a base of 2. The black line represents the average TM-scores under each *Neff* bin with a bin width of two.

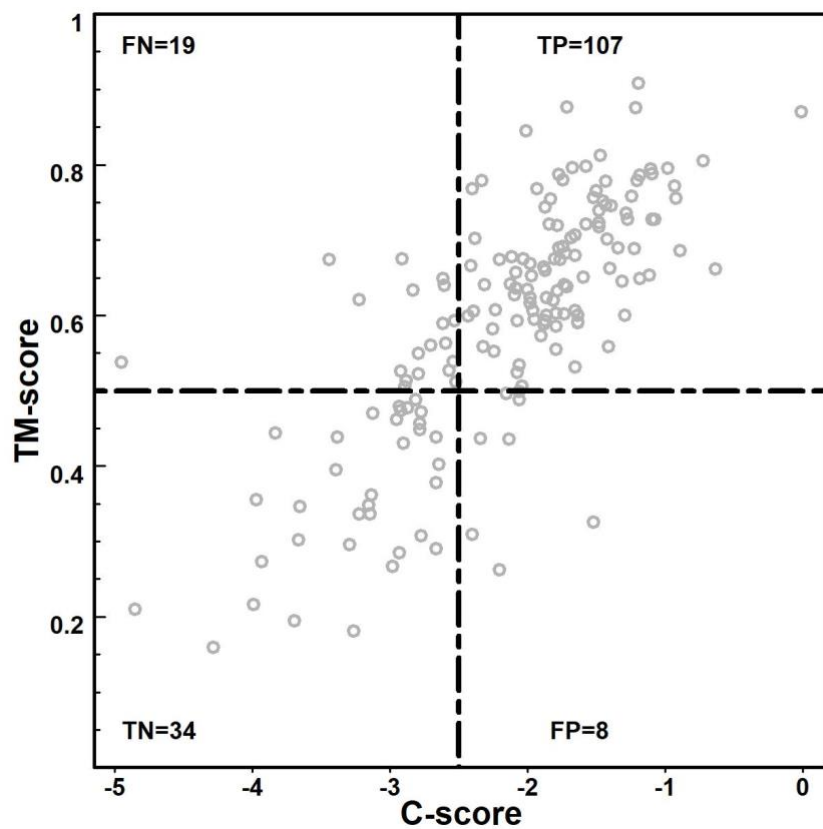
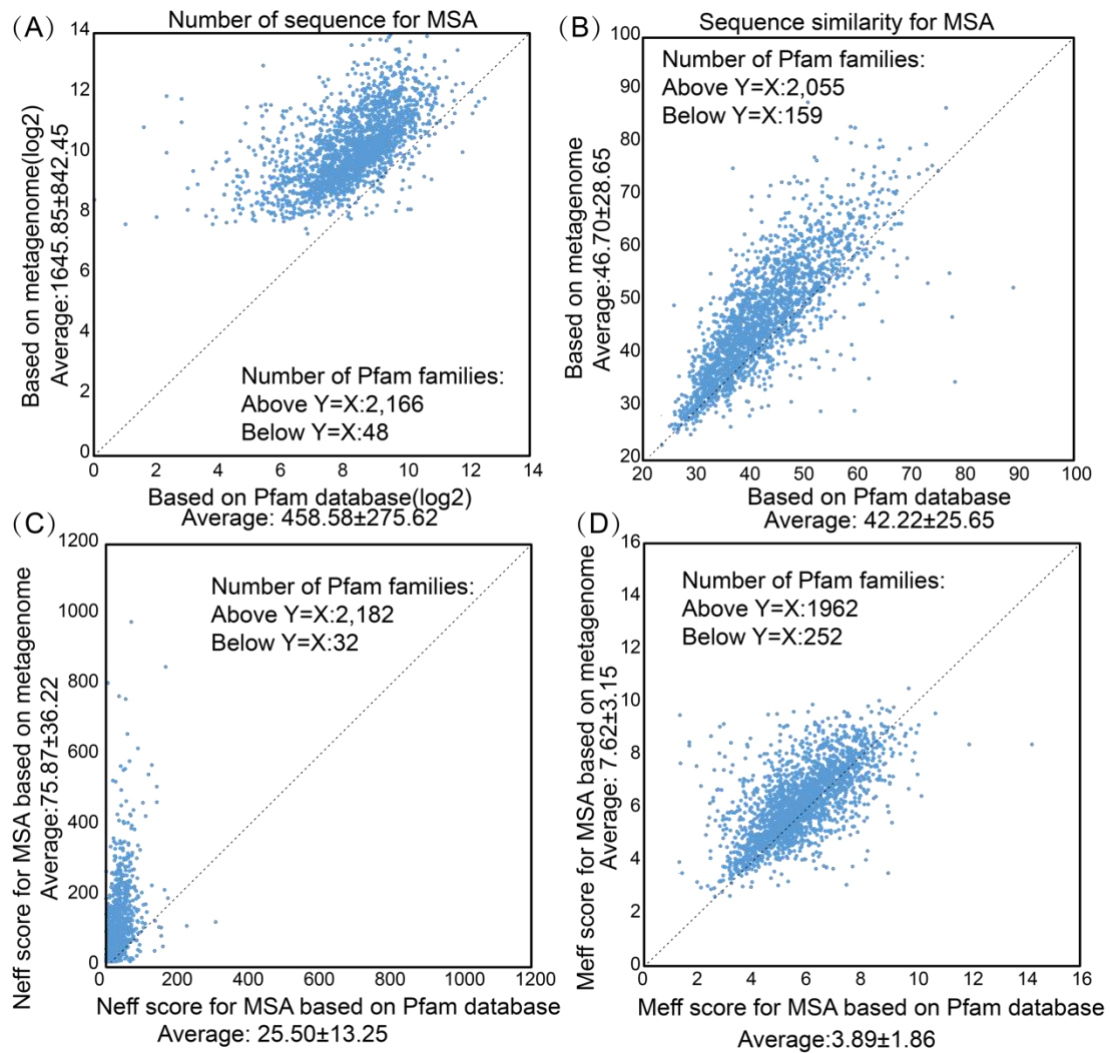
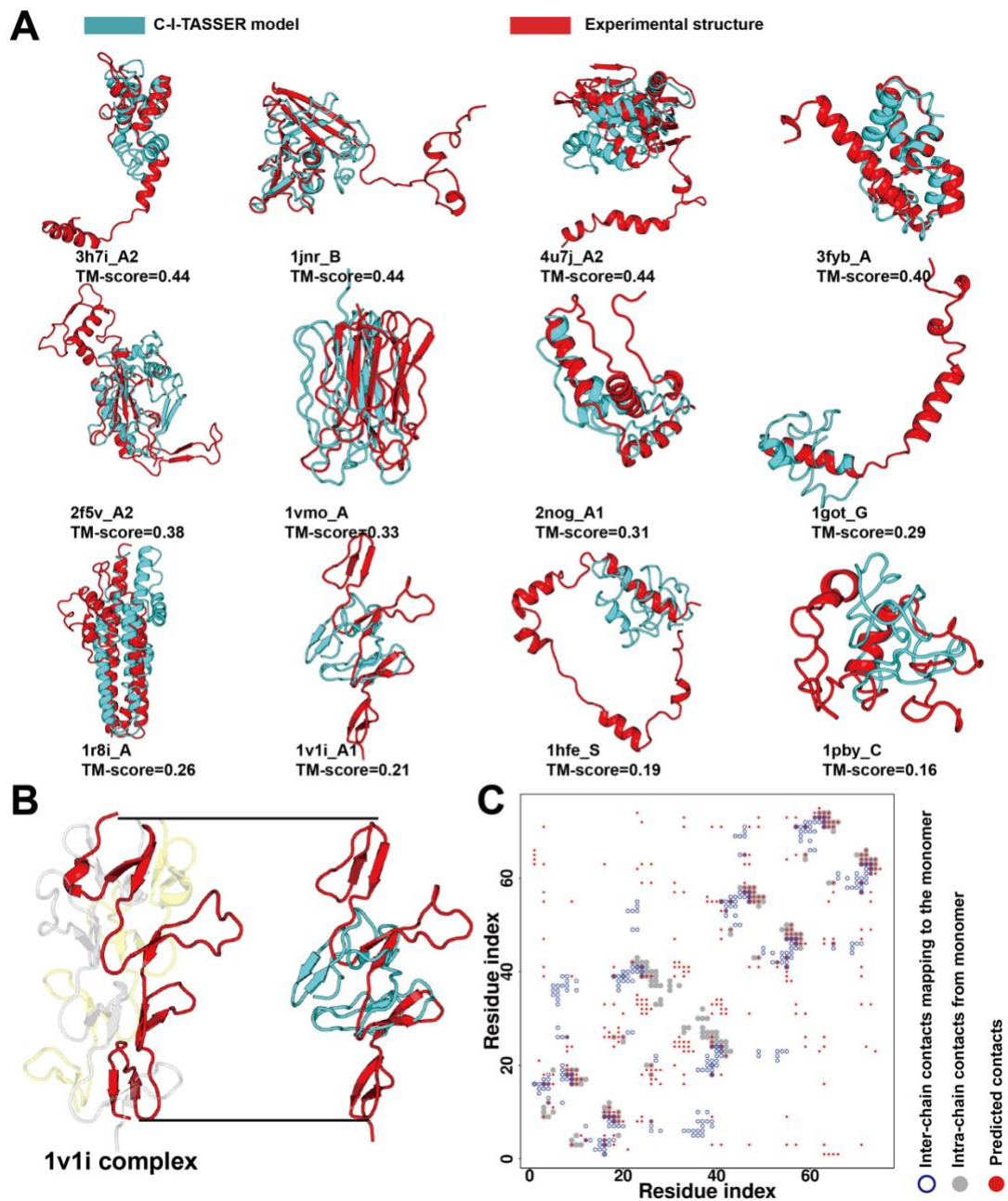


Figure S2. Accuracy estimation of predicted models using C-score defined by Eq. (2), in Materials and Methods. Represented by TM-score of the first C-I-TASSER model versus C-score.

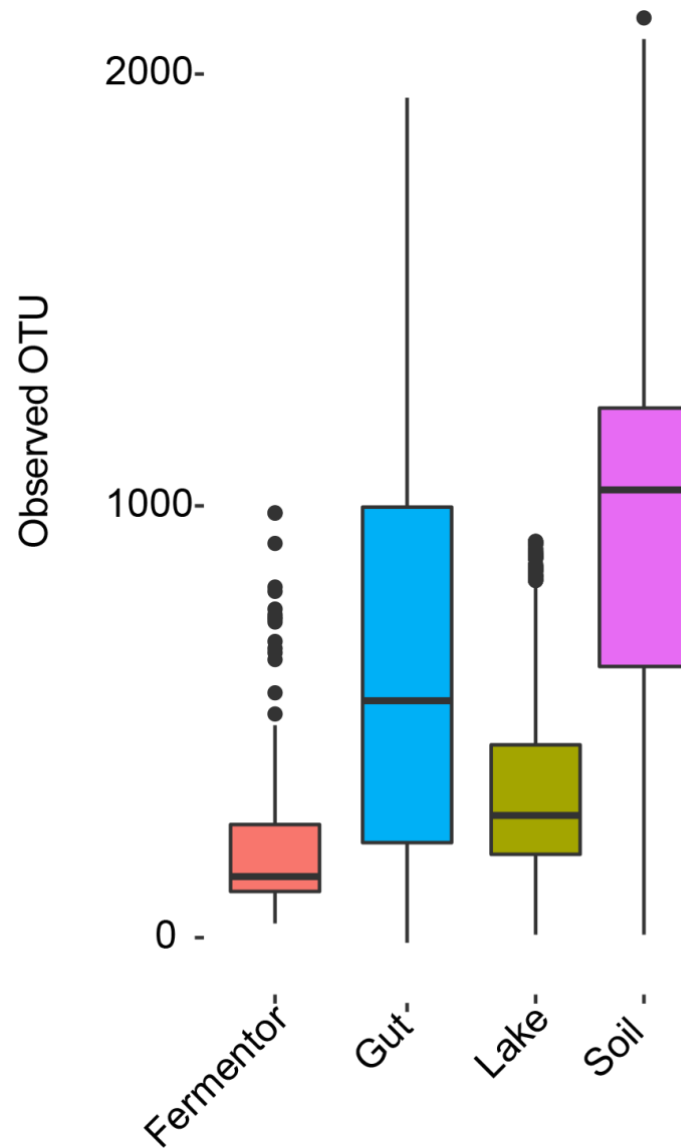


**Figure S3. The comparison of Pfam MSA and Metagenome MSA.** (A) The number of sequences for Pfam families in Pfam database and supplemented by metagenome data. (B) The sequence similarity for MSA of Pfam families to the query in Pfam database and supplemented by metagenome data. (C) The *Neff* score distribution for Pfam families in Pfam database and supplemented by metagenome data. (D) The *Meff* score distribution for Pfam families in Pfam database and supplemented by metagenome data.

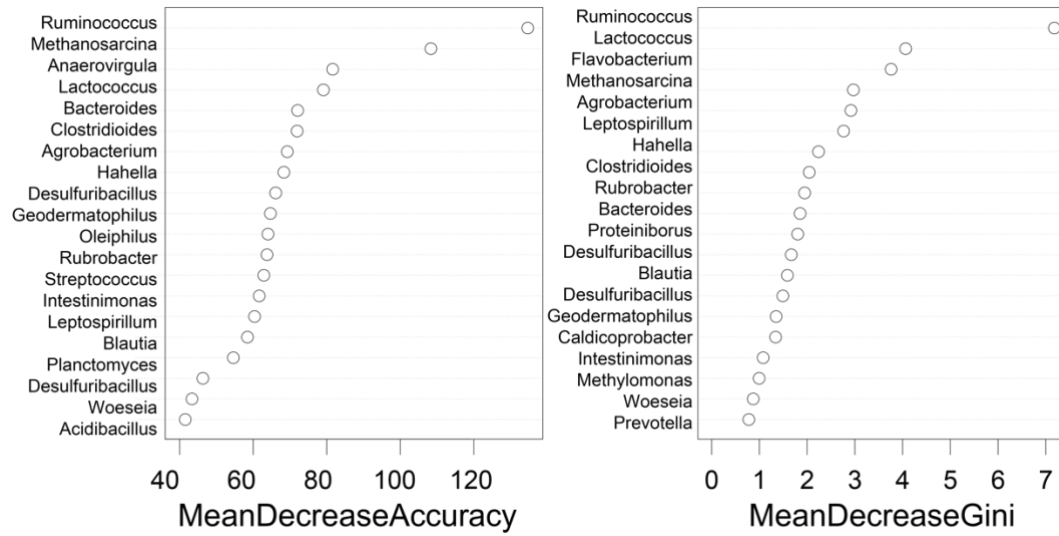




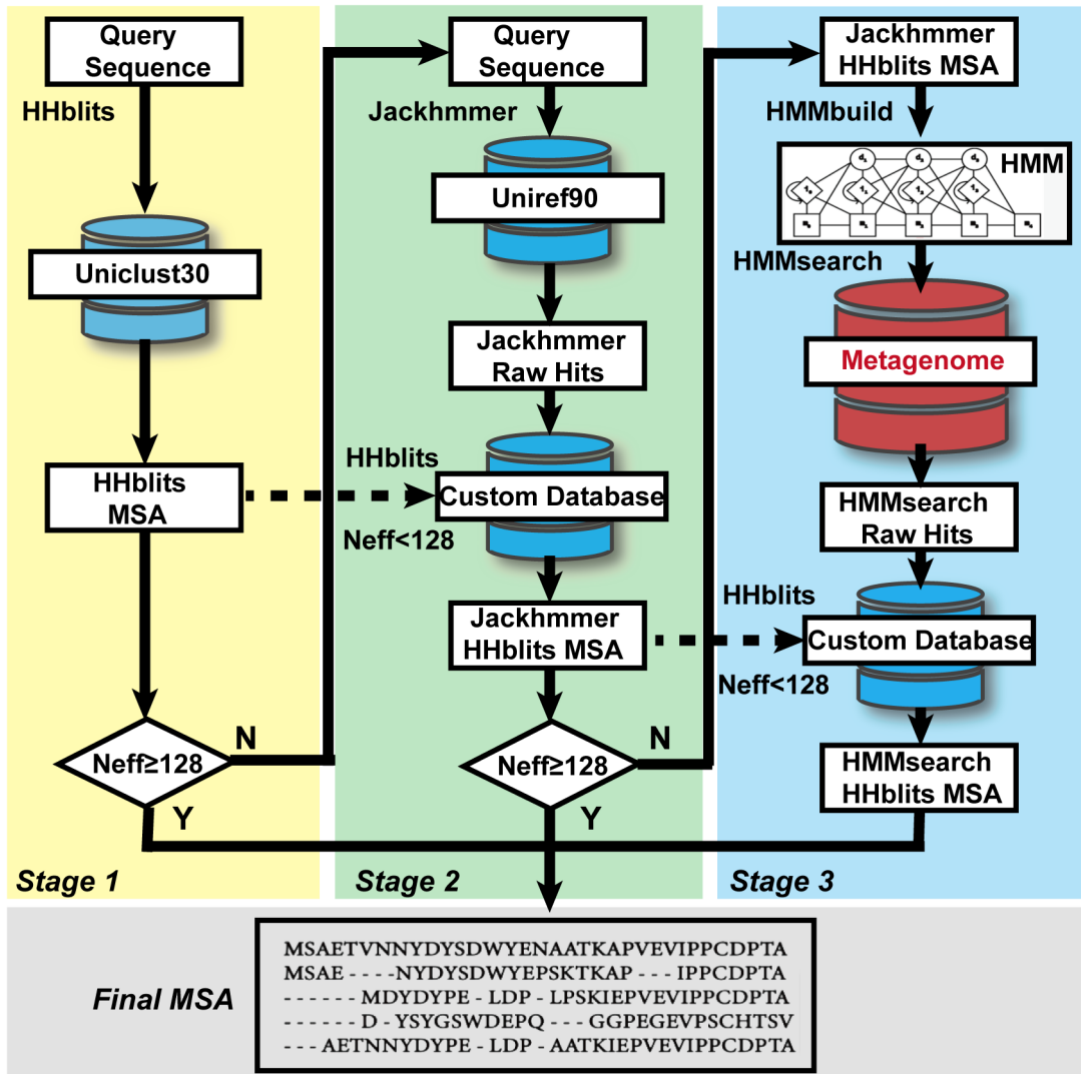
**Figure S4. C-I-TASSER models for 12 cases from 168 proteins benchmark dataset that has large *Neff* but low TM-score.** (A) C-I-TASSER models (cyan) and experimental structures (red) of 12 cases. (B) 1v1i trimer complex (three copies are shown as red, grey and yellow) and C-I-TASSER model (cyan) for the 1v1i\_A1 monomer. (C) Predicted contact map (red) and experimental contact map, where the inter-chain contacts are shown as blue circle and intra-chain contacts are shown as grey points.



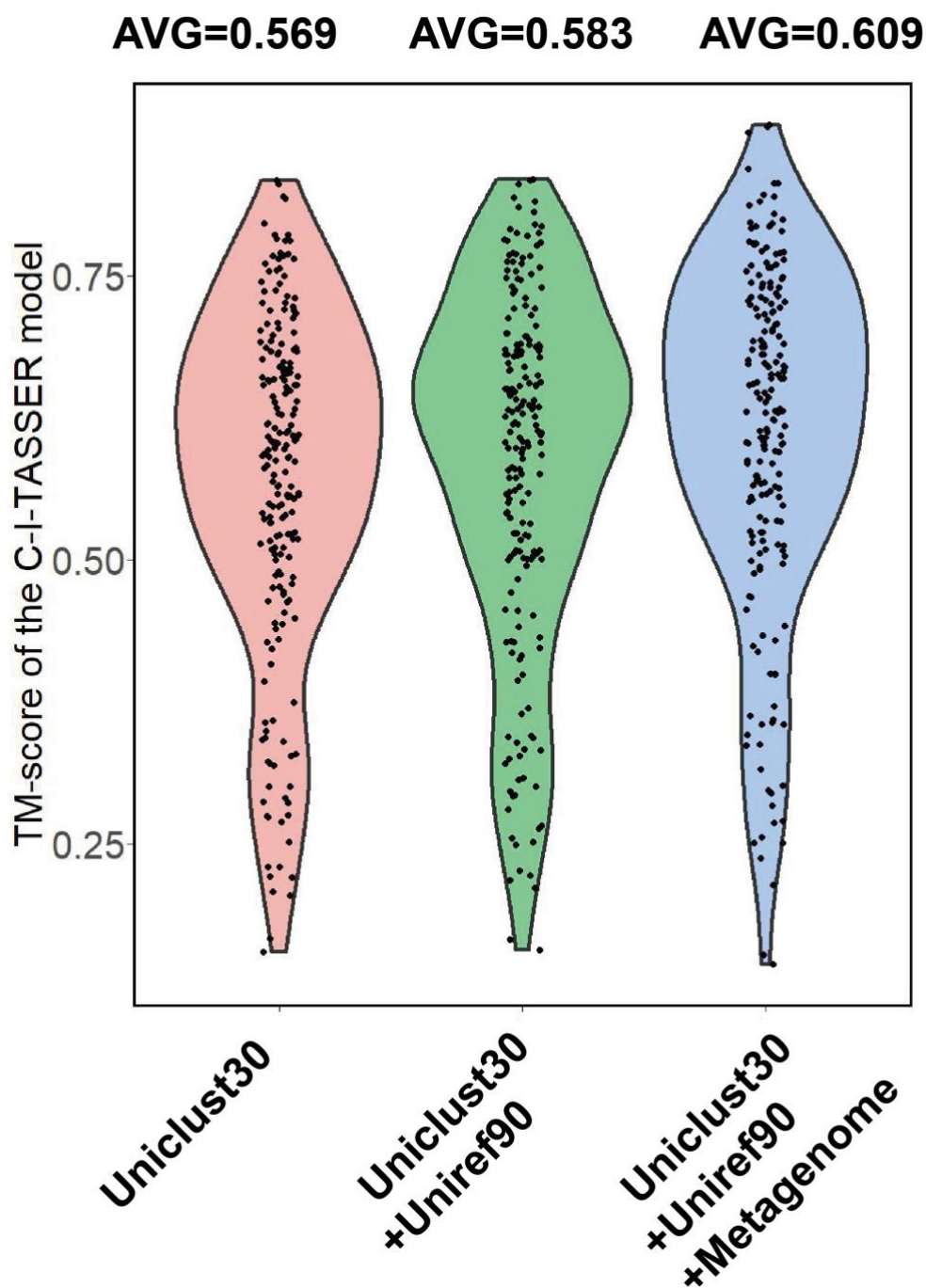
**Figure S5. The species richness statistic for four biomes (Fermentor, Gut, Lake and Soil).** The raw metagenome sequences were assembled, extract the 16s rRNA and clustered by 97% similarity to obtain the operational taxonomic units (OTUs) distribution, sequentially. The OTU distribution could represent the species richness in corresponding biome.



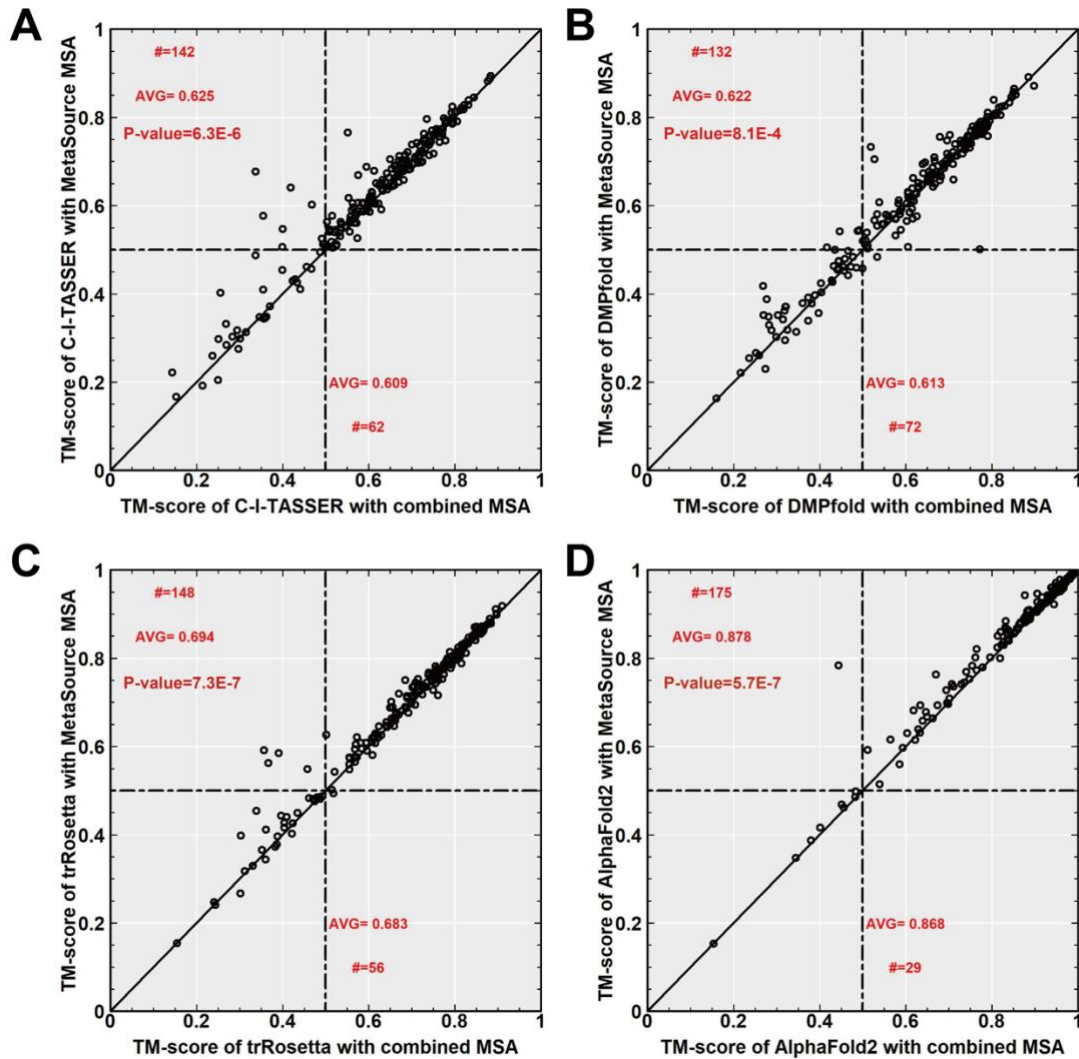
**Figure S6. The top 20 importance features (on genus level) for the multiple-classified Random Forest model.** The importance of features was estimated and ranked by accuracy and Gini index.



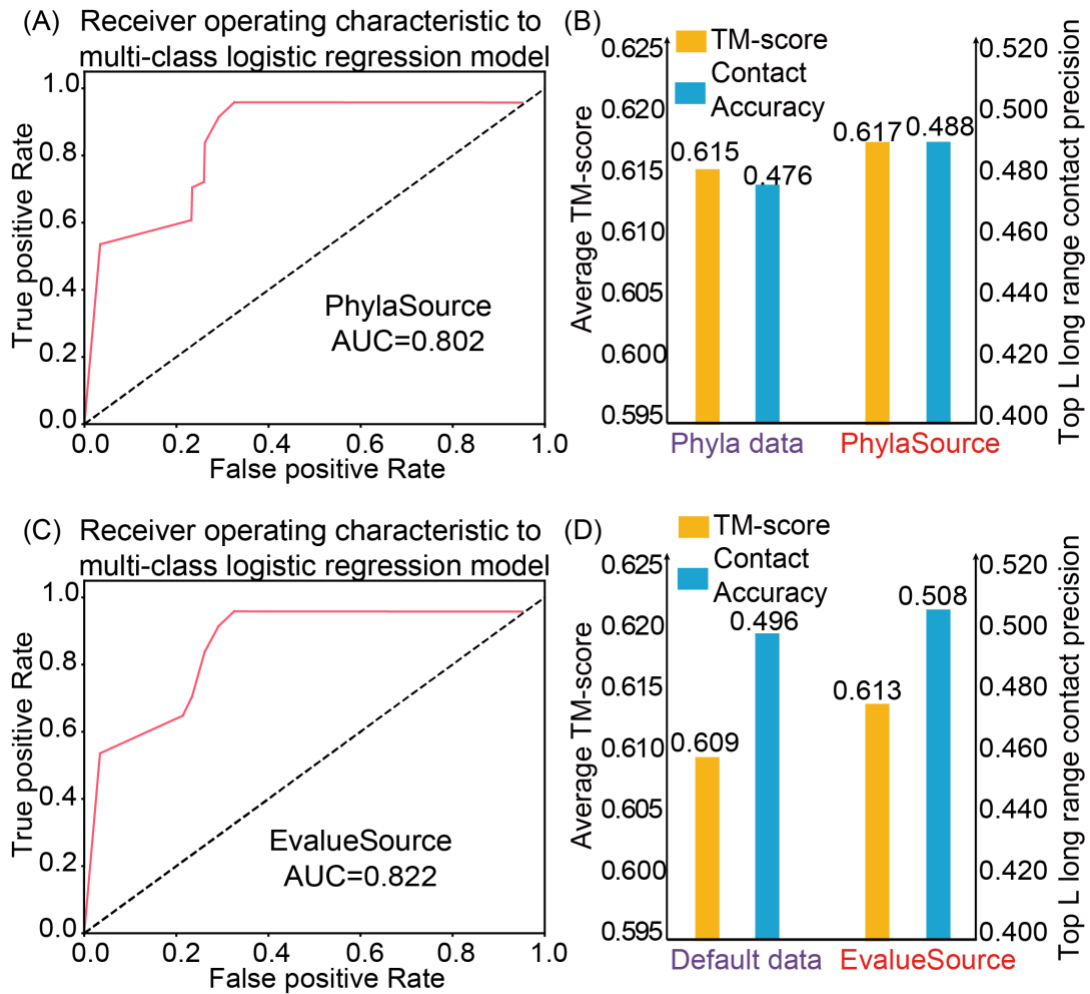
**Figure S7. DeepMSA pipeline for multiple sequence alignment generation.** The metagenome database in the third step can be the combination of four biomes (Fermentor, Gut, Lake and Soil) or each individual biome.



**Figure S8. Modeling results of C-I-TASSER utilizing genome and metagenome databases.** TM-scores of the first model of C-I-TASSER using Uniclust30 (genome) database (A), Uniclust30+Uniref90 (genome) databases (B) and Uniclust30+Uniref90+four biomes metagenome (genome+metagenome) databases (C).

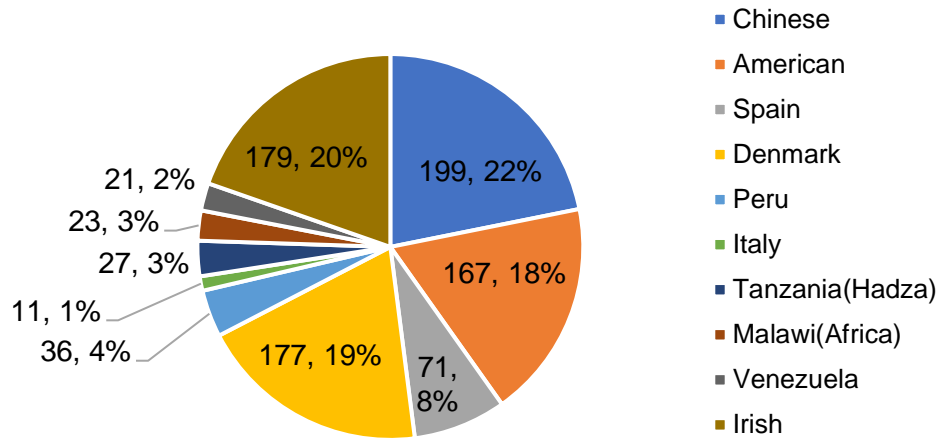


**Figure S9. Head-to-head comparison of the protein folding methods using MetaSource selected biome MSA and combined biome MSA.** TM-score comparison of C-I-TASSER (A), DMPfold (B), trRosetta (C) and AlphaFold2 (D) for the 204 validation Pfam families using MetaSource selected biome MSA and combined biome MSA. P-values are calculated by one-tail paired Student's t-test.



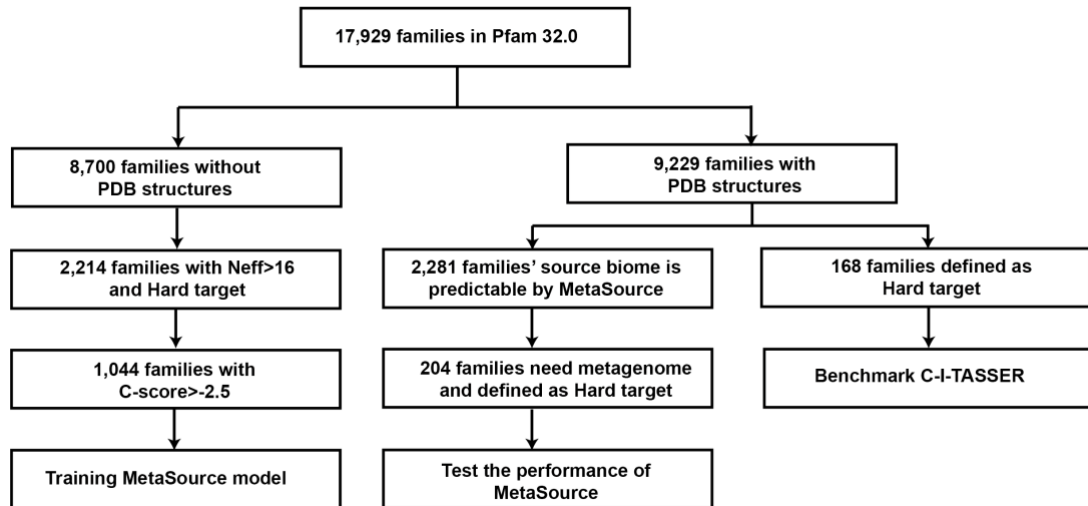
**Figure S10. The construction of the PhylaSource and EvalSource based on 964 Pfam families with unsolved structure.** (A) The Receiver operating characteristic to multi-class logistic regression model. The PhylaSource was constructed by multi-class logistic regression model and the area under curve illustrate that the accuracy of the model is 80.2%. (B) The validation test of PhylaSource. The validation of the PhylaSource was performed by comparing the Pfam families that supplemented by genome data download from NCBI (named as Phyla data) and guided by PhylaSource. The quality of MSA was estimated by TM-score and precision of top-*L* long range contacts. (C) The Receiver operating characteristic to multi-class logistic regression model. The EvalSource was constructed by multi-class logistic regression model and the area under curve illustrate that the accuracy of the model is 82.2%. (D) The validation test of EvalSource. The validation of the EvalSource was performed by comparing the Pfam families that supplemented by metagenome data by DeepMSA with default E-values (named as default data) and guided by EvalSource. The quality of MSA was estimated by TM-score and precision of top-*L* long range contacts.

The number of sequenced individuals  
(number,proportion)

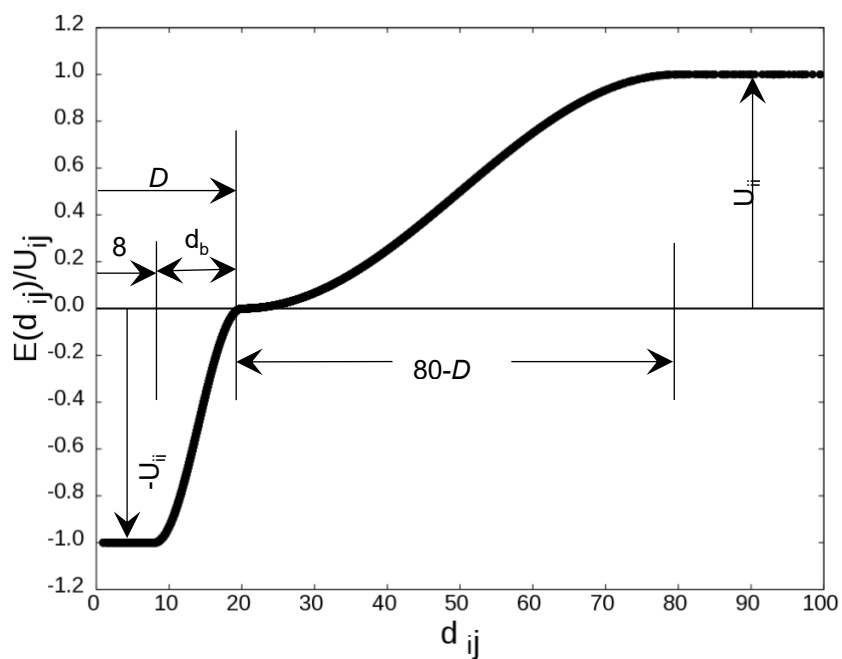


**Figure S11. For the Gut biome, the statistical result based on country distribution.** The 911 samples were collected from 10 countries, covering four continents (Africa, Asia, Europe, Americas).

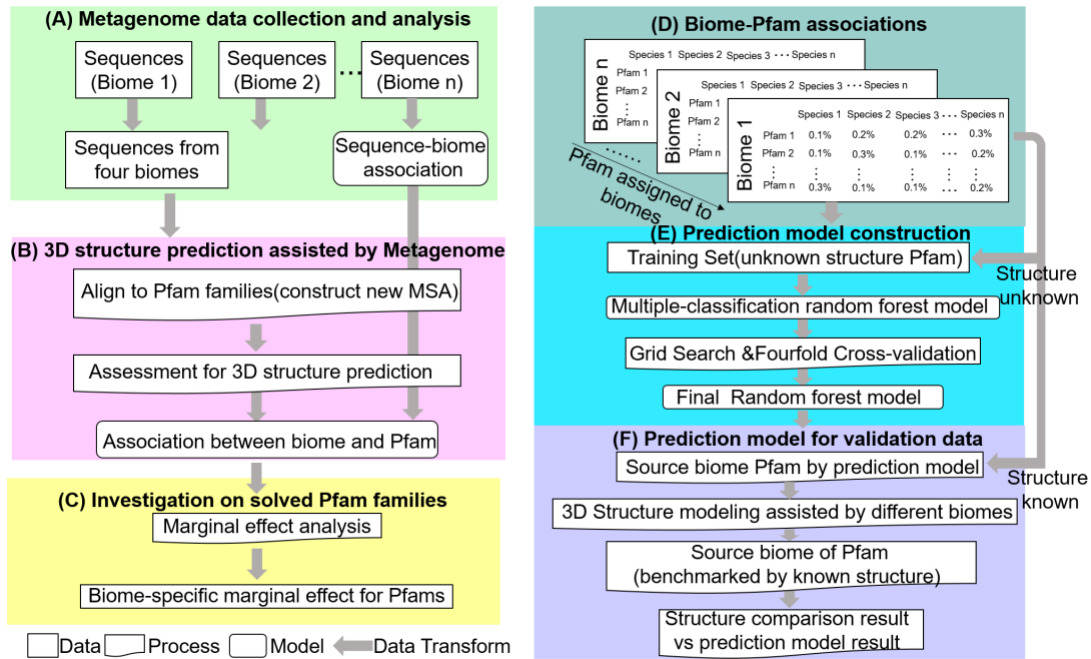




**Figure S12. Data collection flow from Pfam database for training and validating MetaSource and benchmarking the C-I-TASSER.** For 8,700 Pfam families with unsolved structure, 1,044 Pfam families were used to train the MetaSource prediction model after a set of filtration. For 9,229 Pfam families with solved structure has been randomly selected as benchmark dataset to investigate the fold ability of C-I-TASSER, and testing dataset for qualify the performance of MetaSource.



**Figure S13:** A schematic of contact potential,  $E_{contact}(d_{ij})$ , for a contacting residue pair  $i$  and  $j$  as defined in Eq. (S1). Here,  $D$  is the protein length-dependent width of the first well and  $U_{ij}$  is the depth of the energy potential that is proportional to the confidence score of the predicted contact between the residue pair  $i$  and  $j$ .  $d_{ij}$  is the  $C_\beta$  distance between the residue pair.



**Figure S14. Workflow for targeted MetaSource model construction.** (A) Sequences from different biomes were collected, and the biome-sequence associations are also organized. (B) New multiple sequence alignment (MSA) is constructed for Pfam families after search the homology sequences from different biomes. After the MSA is constructed, the *Neff*/C-score were calculated to evaluate the quality of MSA. (C) The marginal effect is evaluated to quantify the effects of metagenome data from different biomes on Pfam families. (D) For each of the Pfam families, the normalized taxonomical composition was used as the feature. The biome with highest *Neff* score was used as the data label after supplementing the homology sequences from four biomes respectively. (E) The multiclass Random-Forest model construction. To find the best combination of model parameters, grid search was applied to exhaustive search over all parameter values and 20 cross-validation iterations. (F) The validation of MetaSource using Pfam families whose structure solved. Assisted by sequences from different biomes, the biome of which the structure that shared most similarity to the known structure is compared with the prediction result of MetaSource.

## Supporting Tables

**Table S1. Wilcox test results for differentiating each pair of two biomes based on species distribution.** Results shown are P-values of the Wilcox test.

	<b>Fermentor</b>	<b>Gut</b>	<b>Lake</b>	<b>Soil</b>
<b>Fermentor</b>		2.53E-08	6.75E-10	8.23E-15
<b>Gut</b>	2.53E-08		4.15E-15	7.23E-18
<b>Lake</b>	6.75E-10	4.15E-15		6.28E-09
<b>Soil</b>	8.23E-15	7.23E-18	6.28E-09	

**Table S2. Summary of C-I-TASSER modeling results for 28 Pfam families which has solved experimental structure.** The comparison results for the solved protein to the C-I-TASSER model using TM-align and calculate the TM-score between the C-I-TASSER model and the map experimental structure.

<b>Target</b>	<b><i>Neff</i> of MSA</b>	<b>PDB</b>	<b>TM-score</b>	<b>C-score</b>
<b>PF04213</b>	106.2	6JSB_A	0.841	0.35
<b>PF09139</b>	34.5	6IG4_B	0.763	-1.16
<b>PF03981</b>	309.7	6RWT_A	0.753	-1.65
<b>PF05914</b>	19.1	6U42_4Q	0.742	-9.96
<b>PF01803</b>	28.6	6S9S_A	0.717	-0.99
<b>PF04031</b>	49.1	6OF2_A	0.716	-1.44
<b>PF11704</b>	22.8	6ULG_L	0.684	-0.57
<b>PF12922</b>	48.6	6QJ3_A	0.672	-1.45
<b>PF18755</b>	261.3	6PBD_B	0.651	-1.74
<b>PF10785</b>	33	6GCS_X	0.639	-0.6
<b>PF03381</b>	86.4	6PSY_E	0.635	-0.1
<b>PF04317</b>	30.5	6NZ4_A	0.607	-3.63
<b>PF14687</b>	24.1	6SGB_F6	0.6	-1.31
<b>PF15096</b>	21	6R0X_E	0.556	-4.19
<b>PF12017</b>	87.8	6P5A_A	0.465	-3.43
<b>PF13864</b>	57.3	6U42_5S	0.464	-1.99
<b>PF12357</b>	70	6KZ8_B	0.401	-2.33
<b>PF14260</b>	101	6P1H_A	0.381	-5.47
<b>PF04281</b>	32.8	6JNF_C	0.357	-3.97
<b>PF14636</b>	16.1	6ULG_N	0.322	-2.51
<b>PF07127</b>	76.8	6U6G_A	0.308	-4.15
<b>PF03963</b>	196.5	6IEE_B	0.272	-3.66
<b>PF12542</b>	45.8	5YZG_X	0.255	-3.58
<b>PF14960</b>	18.4	6J5J_i	0.249	-3.07
<b>PF14892</b>	24.1	6U42_7H	0.239	-3.7
<b>PF10172</b>	20.6	6Q0R_E	0.221	-3.26
<b>PF13868</b>	44.2	6U42_4Y	0.213	-3.8
<b>PF08648</b>	70.8	6QX9_X	0.182	-3.96

**Table S3. The contact precision on the 12 cases in the benchmark dataset that has large  $N_{eff} > 16$  but with low TM-score shown in Figure S1. The nine columns show the top  $L$ ,  $L/2$ , and  $L/5$  contacts as well as the long-, medium- and short-range contacts.**

Target	Short range			Medium range			Long range		
	$L/5$	$L/2$	$L$	$L/5$	$L/2$	$L$	$L/5$	$L/2$	$L$
3h7i_A2	0.400	0.274	0.216	0.200	0.177	0.120	0.760	0.484	0.312
1jnr_B	0.793	0.541	0.302	0.655	0.378	0.302	0.966	0.514	0.315
4u7j_A2	1.000	0.611	0.323	0.911	0.602	0.310	1.000	0.858	0.673
3fyb_A	0.790	0.388	0.225	0.684	0.306	0.153	0.105	0.204	0.214
2f5v_A2	0.590	0.354	0.192	0.769	0.505	0.318	0.949	0.838	0.657
1vmo_A	0.781	0.444	0.227	1.000	0.877	0.540	1.000	0.889	0.755
2nog_A1	0.588	0.250	0.125	0.647	0.296	0.148	0.882	0.796	0.534
1got_G	0.273	0.103	0.052	0.000	0.000	0.000	0.000	0.000	0.000
1r8i_A	0.081	0.075	0.064	0.135	0.054	0.037	0.081	0.054	0.032
1v1i_A1	0.867	0.730	0.533	0.467	0.297	0.213	0.000	0.000	0.000
1hfe_S	0.471	0.273	0.159	0.059	0.023	0.011	0.000	0.000	0.000
1pby_C	0.400	0.333	0.291	0.133	0.077	0.038	0.200	0.128	0.063

**Table S4. The statistical result for GO annotations (level 3) which were only detected in single biome for the 964 Pfam families.** The numbers count for the GO entries that are only detected in a specific biome. The proportion of all entries detected in the corresponding biome under the specific top GO annotation was calculated.

<b>Biome</b>	<b>Biological Process</b>	<b>Molecular Function</b>	<b>Cellular Component</b>
<b>Gut</b>	21(30%)	15(22%)	18(25%)
<b>Lake</b>	18(21%)	11(25%)	17(30%)
<b>Soil</b>	42(33%)	25(25%)	44(30%)
<b>Fermentor</b>	48(35%)	18(25%)	30(33%)

**Table S5. Ten case studies for illustration of the Pfam-biome associations.** Ten Pfam families were selected based on the record in Pfam database and literature review. “Ferm” refers to “Fermentor”; “Data1” to “Uniref100”; “Data2” to “IMG+Uniref100”; “Data3” to “Tara Oceans+Uniref100”; “Data4” to “Metaclust+Uniref100”; “This work” to “Specific biom+Uniref100”. Bold fonts highlight the best result for each target.

Pfam_ID	Source biome	Function	Accuracy	Neff for different databases				
				Data1	Data2	Data3	Data4	This work
<b>PF12652</b>	Ferm	CotJB protein; involed in the synthesis of spore coat related to anaerobic fermentation	Ferm:0.992	59.3	264	95.2	99.6058	<b>Ferm:305.6</b>
<b>PF06135</b>	Gut	IreB regulatory phosphoprotein, cephalosporin resistance	Gut:0.995	37.6	<b>199.3</b>	68.5	90.0926	Gut:187.6
<b>PF07593</b>	Lake	ASPIC and UnbV	Lake:0.961	180.8	881.5	202.5	780.3	<b>Lake:984.4</b>
<b>PF09650</b>	Soil	Putative polyhydroxyalkanoic acid(PHA) system protein, detect in soil,Production of bioplastic	Soil:0.954	36.6	722.1	612.5	309.7	<b>Soil:728.6</b>
<b>PF13822</b>		Acyl-CoA carboxylase epsilon subunit,involved in the biosynthesis of long-chain fatty acids	Soil:0.928	103.7	109.5	160.3	125.6	<b>Soil:309.1</b>
<b>PF04066</b>		Multiple resistance and pH regulation protein F	Soil:0.936	168.1	844.8	240.502	525.5519	<b>Soil:924.0</b>
<b>PF09907</b>		HigB_toxin, RelE-like toxic component of a toxin-antitoxin system	Soil:0.951	80	849.8	313.6	579.4	<b>Soil:927.6</b>
<b>PF09828</b>		Chromate resistance exported protein	Soil:0.968	28.4	633.6	452.5	287.4	<b>Soil:687.9</b>
<b>PF05120</b>		Gas vesicle protein G	Soil:0.986	32	336.8	69	178.9	<b>Soil:487.5</b>
<b>PF05425</b>		Copper resistance protein D	Soil:0.961	257.2	389.5	364.2	726.1	<b>Soil:807.8</b>



**Table S6. The validation result of the MetaSource for the 204 Pfam families with solved structures.** The predicted source biomes by MetaSource are listed together with the biomes that resulted in the higher *Neff* and TM-score for different Pfam families, where accuracies of 79.9% and 80.2% have been achieved by MetaSource on *Neff* and TM-score, respectively. Here, “ferm” refers to “fermentor”.

Pfam	Probability of source biome				Predicted biome	Result based on	
	gut	lake	soil	ferm		<i>Neff</i>	TM-score
PF00284	2.13E-01	6.40E-02	6.44E-01	7.80E-02	soil	soil	soil
PF00631	8.27E-01	6.45E-02	1.32E-02	9.52E-02	gut	gut	gut
PF00647	1.26E-02	8.43E-02	8.22E-01	8.14E-02	soil	soil	soil
PF00658	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
PF00737	6.40E-02	6.44E-01	2.13E-01	7.80E-02	lake	lake	lake
PF00827	1.13E-01	5.39E-02	7.75E-01	5.78E-02	soil	ferm	ferm
PF00833	1.13E-01	1.53E-01	6.91E-01	4.29E-02	soil	lake	lake
PF00838	1.26E-02	8.22E-01	8.43E-02	8.14E-02	lake	lake	lake
PF00853	3.88E-02	8.74E-01	4.22E-02	4.53E-02	lake	lake	lake
PF00960	1.38E-02	1.66E-01	6.57E-01	1.63E-01	soil	soil	soil
PF01049	8.74E-01	4.22E-02	3.88E-02	4.53E-02	gut	gut	gut
PF01111	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	lake	lake
PF01115	1.26E-02	8.43E-02	8.22E-01	8.14E-02	soil	ferm	ferm
PF01125	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
PF01140	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
PF01191	1.13E-01	5.39E-02	7.75E-01	5.78E-02	soil	soil	soil
PF01194	1.09E-01	6.52E-01	1.73E-01	6.60E-02	lake	lake	lake
PF01200	1.09E-01	6.52E-01	1.73E-01	6.60E-02	lake	lake	lake
PF01213	3.22E-02	8.54E-01	6.81E-02	4.52E-02	lake	lake	lake
PF01214	7.46E-01	7.94E-02	1.09E-01	6.60E-02	gut	gut	gut
PF01247	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
PF01267	1.03E-02	8.53E-01	5.58E-02	8.14E-02	lake	soil	lake
PF01278	1.04E-01	2.96E-02	8.04E-01	6.19E-02	soil	soil	soil
PF01320	1.64E-02	7.86E-01	4.23E-02	1.55E-01	lake	lake	lake
PF01340	4.03E-03	1.86E-02	9.51E-01	2.66E-02	soil	soil	soil
PF01356	1.16E-01	6.08E-01	1.09E-01	1.67E-01	lake	lake	lake
PF01603	7.20E-01	8.82E-02	1.12E-01	7.99E-02	gut	gut	gut
PF01716	2.13E-01	6.40E-02	6.44E-01	7.80E-02	soil	soil	soil
PF01780	1.09E-01	6.52E-01	1.73E-01	6.60E-02	lake	lake	lake
PF01793	9.01E-01	2.57E-02	2.69E-02	4.62E-02	gut	gut	gut
PF01815	1.61E-02	7.82E-02	8.17E-01	8.87E-02	soil	soil	soil
PF01821	5.13E-01	7.78E-02	3.58E-01	5.13E-02	gut	gut	gut
PF01828	8.85E-02	4.91E-02	1.77E-01	6.85E-01	ferm	ferm	ferm
PF01893	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
PF01993	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	soil	soil
PF02015	2.99E-02	7.56E-01	5.93E-02	1.55E-01	lake	lake	lake
PF02064	1.08E-02	9.20E-01	2.95E-02	3.97E-02	lake	lake	lake
PF02093	3.58E-01	7.78E-02	5.13E-01	5.13E-02	soil	ferm	ferm
PF02100	9.20E-01	2.95E-02	1.08E-02	3.97E-02	gut	gut	gut

<b>PF02145</b>	1.37E-02	8.13E-01	6.95E-02	1.04E-01	lake	lake	lake
<b>PF02177</b>	8.74E-01	4.22E-02	3.88E-02	4.53E-02	gut	gut	gut
<b>PF02209</b>	1.33E-02	8.52E-01	5.15E-02	8.34E-02	lake	lake	lake
<b>PF02240</b>	2.06E-02	8.27E-02	1.16E-01	7.81E-01	ferm	ferm	ferm
<b>PF02253</b>	0.00E+00	4.13E-03	9.57E-01	3.92E-02	soil	ferm	ferm
<b>PF02271</b>	1.07E-01	5.09E-02	7.77E-01	6.60E-02	soil	ferm	ferm
<b>PF02284</b>	9.20E-01	2.95E-02	1.08E-02	3.97E-02	gut	gut	gut
<b>PF02289</b>	2.06E-02	2.94E-01	4.90E-01	1.96E-01	soil	soil	soil
<b>PF02312</b>	3.88E-02	4.22E-02	8.74E-01	4.53E-02	soil	soil	soil
<b>PF02315</b>	1.93E-02	3.04E-01	6.04E-02	6.17E-01	ferm	soil	soil
<b>PF02531</b>	6.44E-01	6.40E-02	2.13E-01	7.80E-02	gut	gut	gut
<b>PF02605</b>	2.13E-01	6.44E-01	6.40E-02	7.80E-02	lake	lake	lake
<b>PF02611</b>	1.07E-01	2.62E-02	6.41E-01	2.26E-01	soil	ferm	ferm
<b>PF02679</b>	6.81E-02	8.30E-01	2.24E-02	7.93E-02	lake	lake	lake
<b>PF02792</b>	1.09E-01	7.46E-01	7.94E-02	6.60E-02	lake	lake	lake
<b>PF02840</b>	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
<b>PF02888</b>	7.61E-01	7.96E-02	9.27E-02	6.68E-02	gut	gut	gut
<b>PF02898</b>	2.15E-01	4.74E-02	5.91E-01	1.46E-01	soil	soil	soil
<b>PF02921</b>	1.12E-01	8.82E-02	7.20E-01	7.99E-02	soil	ferm	ferm
<b>PF02924</b>	0.00E+00	4.13E-03	8.71E-01	1.25E-01	soil	lake	ferm
<b>PF02963</b>	1.42E-02	8.32E-01	3.96E-02	1.14E-01	lake	lake	lake
<b>PF02974</b>	7.94E-01	5.60E-02	9.33E-03	1.40E-01	gut	gut	gut
<b>PF02975</b>	8.27E-03	3.58E-02	8.96E-01	6.04E-02	soil	ferm	ferm
<b>PF02979</b>	2.12E-01	7.23E-01	4.02E-02	2.40E-02	lake	lake	lake
<b>PF03013</b>	6.51E-01	3.07E-01	4.49E-03	3.71E-02	gut	gut	gut
<b>PF03095</b>	7.46E-01	7.94E-02	1.09E-01	6.60E-02	gut	gut	gut
<b>PF03110</b>	2.13E-01	6.40E-02	6.44E-01	7.80E-02	soil	soil	soil
<b>PF03126</b>	1.09E-01	7.46E-01	7.94E-02	6.60E-02	lake	lake	lake
<b>PF03288</b>	3.63E-01	1.60E-02	4.77E-01	1.44E-01	soil	soil	soil
<b>PF03411</b>	1.09E-01	1.47E-01	6.25E-01	1.18E-01	soil	ferm	ferm
<b>PF03416</b>	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	ferm	ferm
<b>PF03502</b>	9.51E-01	1.86E-02	4.03E-03	2.66E-02	gut	gut	gut
<b>PF03660</b>	1.97E-01	7.32E-02	6.64E-01	6.60E-02	soil	ferm	ferm
<b>PF03735</b>	1.10E-01	4.66E-02	7.76E-01	6.80E-02	soil	soil	soil
<b>PF03829</b>	2.54E-03	1.12E-01	7.61E-01	1.25E-01	soil	soil	soil
<b>PF03870</b>	1.09E-01	7.46E-01	7.94E-02	6.60E-02	lake	lake	lake
<b>PF03887</b>	7.13E-03	7.13E-01	3.47E-02	2.45E-01	lake	lake	lake
<b>PF03925</b>	1.08E-02	4.09E-02	8.74E-01	7.48E-02	soil	soil	soil
<b>PF03974</b>	7.74E-01	4.09E-02	1.08E-02	1.75E-01	gut	gut	gut
<b>PF03997</b>	1.09E-01	7.46E-01	7.94E-02	6.60E-02	lake	lake	lake
<b>PF04008</b>	1.03E-01	8.11E-01	1.38E-02	7.22E-02	lake	lake	lake
<b>PF04038</b>	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	soil	soil
<b>PF04062</b>	1.07E-01	7.77E-01	5.09E-02	6.60E-02	lake	lake	lake
<b>PF04098</b>	0.00E+00	4.58E-02	9.04E-01	5.00E-02	soil	soil	soil
<b>PF04216</b>	1.00E-01	1.03E-01	4.87E-01	3.10E-01	soil	soil	soil
<b>PF04269</b>	1.08E-02	8.74E-01	4.09E-02	7.48E-02	lake	lake	lake

<b>PF04270</b>	7.53E-01	1.67E-01	1.61E-02	6.37E-02	gut	gut	gut
<b>PF04300</b>	7.59E-01	4.95E-02	1.03E-02	1.81E-01	gut	gut	gut
<b>PF04362</b>	3.97E-02	3.17E-01	5.93E-02	5.84E-01	gut	ferm	ferm
<b>PF04386</b>	1.04E-01	1.17E-01	1.18E-02	7.67E-01	gut	ferm	ferm
<b>PF04433</b>	7.20E-01	8.82E-02	1.12E-01	7.99E-02	gut	gut	gut
<b>PF04502</b>	1.09E-01	7.94E-02	6.60E-02	7.46E-01	gut	ferm	ferm
<b>PF04591</b>	8.17E-01	7.82E-02	1.61E-02	8.87E-02	gut	gut	gut
<b>PF04621</b>	3.55E-01	5.20E-01	7.58E-02	4.93E-02	lake	lake	lake
<b>PF04721</b>	3.88E-02	8.74E-01	4.22E-02	4.53E-02	lake	lake	lake
<b>PF04729</b>	1.09E-01	7.94E-02	6.60E-02	7.46E-01	gut	ferm	ferm
<b>PF04739</b>	1.09E-01	7.94E-02	6.60E-02	7.46E-01	gut	ferm	ferm
<b>PF05005</b>	1.35E-01	7.98E-01	3.73E-02	2.99E-02	lake	lake	lake
<b>PF05023</b>	8.52E-03	2.31E-02	8.15E-01	1.53E-01	soil	soil	soil
<b>PF05026</b>	7.51E-01	5.96E-02	1.09E-01	7.99E-02	gut	gut	gut
<b>PF05153</b>	1.87E-01	2.45E-02	1.60E-01	6.29E-01	gut	ferm	ferm
<b>PF05247</b>	8.27E-03	6.87E-01	2.44E-01	6.04E-02	lake	lake	lake
<b>PF05280</b>	1.08E-02	6.74E-01	2.41E-01	7.48E-02	lake	lake	lake
<b>PF05303</b>	1.03E-02	8.59E-01	4.95E-02	8.14E-02	lake	lake	lake
<b>PF05321</b>	1.61E-02	8.17E-01	7.82E-02	8.87E-02	lake	lake	lake
<b>PF05354</b>	2.13E-01	6.55E-01	6.07E-02	7.18E-02	lake	lake	lake
<b>PF05370</b>	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
<b>PF05551</b>	2.16E-01	6.54E-01	9.57E-02	3.49E-02	lake	lake	lake
<b>PF05854</b>	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
<b>PF05856</b>	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
<b>PF05870</b>	8.47E-01	1.73E-02	3.10E-03	1.33E-01	gut	gut	gut
<b>PF05983</b>	1.12E-01	7.20E-01	8.82E-02	7.99E-02	lake	lake	lake
<b>PF06141</b>	2.29E-01	5.30E-01	1.55E-01	8.62E-02	lake	lake	lake
<b>PF06154</b>	1.61E-02	8.17E-01	7.82E-02	8.87E-02	lake	lake	lake
<b>PF06175</b>	1.11E-01	1.34E-01	6.78E-02	6.88E-01	gut	ferm	ferm
<b>PF06304</b>	2.10E-01	2.09E-01	5.67E-01	1.39E-02	soil	soil	soil
<b>PF06384</b>	6.67E-01	4.96E-02	1.07E-01	1.76E-01	gut	gut	gut
<b>PF06400</b>	3.58E-02	4.02E-02	8.81E-01	4.33E-02	soil	soil	soil
<b>PF06438</b>	1.08E-02	4.62E-02	9.06E-02	8.52E-01	gut	ferm	ferm
<b>PF06456</b>	3.88E-02	4.22E-02	4.53E-02	8.74E-01	ferm	ferm	ferm
<b>PF06475</b>	2.12E-01	7.06E-01	4.75E-02	3.42E-02	lake	lake	lake
<b>PF06482</b>	3.88E-02	4.22E-02	4.53E-02	8.74E-01	ferm	ferm	ferm
<b>PF06557</b>	2.06E-02	7.81E-01	8.27E-02	1.16E-01	lake	lake	lake
<b>PF06684</b>	3.67E-01	2.17E-01	2.04E-01	2.12E-01	gut	gut	gut
<b>PF06844</b>	1.04E-01	1.19E-01	7.41E-01	3.64E-02	soil	ferm	ferm
<b>PF06870</b>	7.51E-01	5.96E-02	1.09E-01	7.99E-02	gut	gut	gut
<b>PF07072</b>	1.08E-02	3.40E-01	5.90E-01	6.00E-02	soil	ferm	ferm
<b>PF07152</b>	1.04E-01	6.67E-01	1.17E-01	1.12E-01	lake	lake	lake
<b>PF07262</b>	1.99E-02	8.22E-01	8.37E-02	7.43E-02	lake	lake	lake
<b>PF07352</b>	2.38E-03	5.29E-01	4.12E-01	5.71E-02	lake	ferm	ferm
<b>PF07361</b>	1.08E-02	3.97E-02	8.90E-01	6.00E-02	soil	soil	soil
<b>PF07408</b>	6.69E-01	5.97E-02	1.52E-01	1.19E-01	gut	gut	gut

<b>PF07460</b>	1.19E-01	5.82E-02	5.74E-01	2.48E-01	soil	soil	soil
<b>PF07472</b>	1.82E-02	5.41E-02	1.16E-01	8.12E-01	gut	ferm	ferm
<b>PF07682</b>	2.32E-01	5.97E-02	5.89E-01	1.19E-01	soil	soil	soil
<b>PF07828</b>	1.61E-02	7.82E-02	8.17E-01	8.87E-02	soil	soil	soil
<b>PF08000</b>	2.04E-01	4.49E-03	2.41E-01	5.50E-01	gut	ferm	ferm
<b>PF08127</b>	8.91E-02	7.82E-01	7.26E-02	5.64E-02	lake	lake	lake
<b>PF08208</b>	2.98E-02	3.95E-02	4.52E-02	8.85E-01	gut	ferm	ferm
<b>PF08536</b>	2.13E-01	6.40E-02	6.44E-01	7.80E-02	soil	soil	soil
<b>PF08714</b>	9.51E-03	2.60E-01	4.46E-01	2.84E-01	soil	soil	soil
<b>PF08773</b>	2.20E-01	3.84E-02	2.97E-02	7.12E-01	gut	ferm	ferm
<b>PF08804</b>	1.08E-02	8.74E-01	4.09E-02	7.48E-02	lake	lake	lake
<b>PF08814</b>	2.06E-02	8.27E-02	1.16E-01	7.81E-01	gut	ferm	ferm
<b>PF08854</b>	6.31E-01	6.90E-02	2.14E-01	8.64E-02	gut	gut	gut
<b>PF08869</b>	6.29E-01	1.07E-01	2.80E-02	2.36E-01	gut	gut	gut
<b>PF08883</b>	3.20E-02	3.52E-02	8.51E-01	8.14E-02	soil	soil	soil
<b>PF08931</b>	2.06E-02	7.81E-01	8.27E-02	1.16E-01	lake	lake	lake
<b>PF08941</b>	3.58E-02	4.02E-02	8.81E-01	4.33E-02	soil	soil	soil
<b>PF08958</b>	5.89E-01	5.97E-02	2.32E-01	1.19E-01	gut	gut	gut
<b>PF08963</b>	5.89E-01	5.97E-02	2.32E-01	1.19E-01	gut	gut	gut
<b>PF08968</b>	5.89E-01	5.97E-02	2.32E-01	1.19E-01	gut	gut	gut
<b>PF08974</b>	1.09E-01	5.47E-02	1.26E-01	7.10E-01	gut	ferm	ferm
<b>PF08992</b>	8.57E-02	7.22E-02	1.47E-01	6.95E-01	ferm	ferm	ferm
<b>PF09001</b>	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
<b>PF09009</b>	3.20E-02	3.81E-02	1.95E-01	7.35E-01	ferm	ferm	ferm
<b>PF09015</b>	9.79E-03	4.47E-02	5.09E-02	8.95E-01	ferm	ferm	ferm
<b>PF09021</b>	2.13E-01	1.52E-01	1.60E-01	4.74E-01	ferm	ferm	ferm
<b>PF09028</b>	1.82E-02	7.12E-01	1.54E-01	1.16E-01	lake	lake	lake
<b>PF09044</b>	1.37E-02	4.67E-02	8.61E-01	7.85E-02	soil	soil	soil
<b>PF09056</b>	1.49E-02	6.23E-02	7.96E-01	1.27E-01	soil	soil	soil
<b>PF09059</b>	1.08E-02	4.09E-02	8.74E-01	7.48E-02	soil	soil	soil
<b>PF09078</b>	1.08E-02	2.41E-01	7.48E-02	6.74E-01	ferm	ferm	ferm
<b>PF09082</b>	2.06E-02	8.15E-02	2.01E-01	6.97E-01	ferm	ferm	ferm
<b>PF09143</b>	1.37E-02	4.96E-02	8.45E-01	9.18E-02	soil	soil	soil
<b>PF09160</b>	1.61E-02	7.82E-02	8.17E-01	8.87E-02	soil	soil	soil
<b>PF09194</b>	1.82E-02	5.41E-02	8.09E-01	1.19E-01	soil	soil	soil
<b>PF09203</b>	8.69E-01	6.43E-02	1.65E-02	5.06E-02	gut	gut	gut
<b>PF09204</b>	8.12E-01	3.46E-02	8.27E-03	1.46E-01	gut	gut	gut
<b>PF09208</b>	1.47E-01	7.25E-01	2.24E-02	1.06E-01	lake	lake	lake
<b>PF09218</b>	7.81E-01	8.27E-02	2.06E-02	1.16E-01	gut	gut	gut
<b>PF09221</b>	9.66E-02	5.97E-02	7.47E-01	9.71E-02	soil	lake	soil
<b>PF09223</b>	1.35E-01	1.89E-01	6.04E-01	7.18E-02	soil	soil	soil
<b>PF09225</b>	2.10E-01	1.06E-02	4.55E-01	3.25E-01	soil	soil	soil
<b>PF09226</b>	1.82E-02	5.41E-02	8.09E-01	1.19E-01	soil	soil	soil
<b>PF09233</b>	8.51E-01	5.43E-02	9.25E-03	8.51E-02	gut	gut	gut
<b>PF09391</b>	8.25E-01	4.80E-02	1.03E-01	2.46E-02	gut	gut	gut
<b>PF09392</b>	9.51E-01	1.86E-02	4.03E-03	2.66E-02	gut	gut	gut

<b>PF09393</b>	2.13E-01	1.60E-01	5.14E-01	1.13E-01	soil	soil	soil
<b>PF09412</b>	3.83E-02	3.72E-02	3.69E-02	8.88E-01	ferm	ferm	ferm
<b>PF09449</b>	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	gut	gut
<b>PF09628</b>	2.32E-01	5.89E-01	5.97E-02	1.19E-01	lake	lake	lake
<b>PF09642</b>	2.32E-01	5.89E-01	5.97E-02	1.19E-01	lake	lake	lake
<b>PF10054</b>	1.09E-01	8.22E-01	4.93E-02	2.03E-02	lake	lake	lake
<b>PF10120</b>	1.11E-01	7.34E-01	6.89E-02	8.57E-02	lake	lake	lake
<b>PF10634</b>	2.54E-03	3.82E-01	6.80E-02	5.48E-01	ferm	ferm	ferm
<b>PF11102</b>	9.48E-01	1.86E-02	4.03E-03	2.97E-02	gut	ferm	ferm
<b>PF11419</b>	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	soil	soil
<b>PF11428</b>	2.32E-01	5.97E-02	5.89E-01	1.19E-01	soil	soil	soil
<b>PF11429</b>	2.15E-01	3.40E-02	5.84E-01	1.67E-01	soil	gut	gut
<b>PF11432</b>	2.06E-02	7.81E-01	8.27E-02	1.16E-01	lake	lake	lake
<b>PF11436</b>	2.32E-01	5.97E-02	5.89E-01	1.19E-01	soil	gut	gut
<b>PF11497</b>	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	soil	soil
<b>PF11644</b>	2.06E-02	7.81E-01	8.27E-02	1.16E-01	lake	lake	lake
<b>PF11708</b>	1.12E-01	8.82E-02	7.20E-01	7.99E-02	soil	gut	gut
<b>PF11724</b>	3.27E-01	1.51E-02	5.63E-01	9.54E-02	soil	soil	soil
<b>PF12106</b>	2.38E-01	5.63E-01	9.38E-02	1.05E-01	lake	lake	lake
<b>PF12134</b>	1.09E-01	7.94E-02	7.46E-01	6.60E-02	soil	soil	soil
<b>PF12924</b>	3.88E-02	4.22E-02	8.74E-01	4.53E-02	soil	soil	soil
<b>PF14511</b>	1.68E-02	1.32E-01	7.70E-01	8.07E-02	soil	gut	gut
<b>PF14562</b>	2.06E-02	8.27E-02	7.81E-01	1.16E-01	soil	ferm	ferm
<b>PF15009</b>	2.20E-01	3.84E-02	7.12E-01	2.97E-02	soil	gut	gut
<b>PF18484</b>	1.08E-02	3.41E-01	5.74E-01	7.48E-02	soil	gut	gut
<b>PF18681</b>	2.32E-01	5.97E-02	5.89E-01	1.19E-01	soil	soil	soil
<b>PF18882</b>	1.11E-01	6.89E-02	7.34E-01	8.57E-02	soil	soil	soil

---

**Table S7. Reasons of the C-I-TASSER generated un-foldable model for 29 cases of 204 Pfam validation dataset.** “Combined” means MSA used in C-I-TASSER are generated from the combined four biomes, “MetaSource” means MSA used in C-I-TASSER are generated from the MetaSource selected biome. “Note” column shows the reason why C-I-TASSER failed with this target.

Pfam	TM-score		<i>Neff</i>		Note
	Combined	MetaSource	Combined	MetaSource	
PF00284	0.337	0.487	15.9	9.2	Flexible region in experimental structure
PF00631	0.301	0.298	100.0	97.5	Flexible region in experimental structure
PF01049	0.143	0.221	95.0	93.9	Flexible region in experimental structure
PF01340	0.238	0.259	4.4	3.0	Low <i>Neff</i> and flexible region in experimental structure
PF01780	0.399	0.454	131.1	114.9	Bad N/C terminal orientation of C-I-TASSER model in N/C terminal due to sparse MSA
PF02240	0.355	0.409	4.1	3.9	Low <i>Neff</i>
PF02315	0.297	0.275	6.9	5.5	Low <i>Neff</i> , flexible region in experimental structure and sparse MSA in local region
PF02888	0.152	0.167	20.6	20.2	Flexible region in experimental structure
PF02921	0.250	0.204	135.9	122.1	Sparse MSA in local region
PF02975	0.359	0.346	25.9	23.0	Flexible region in experimental structure
PF03110	0.255	0.402	61.4	57.7	Flexible region in experimental structure
PF03660	0.214	0.192	17.5	15.4	Flexible region in experimental structure
PF04739	0.356	0.344	44.2	41.7	Flexible region in experimental structure
PF05354	0.456	0.461	108.6	92.8	Flexible region in experimental structure
PF05370	0.429	0.433	6.4	5.0	Low <i>Neff</i> and sparse MSA in local region
PF05551	0.371	0.372	168.3	31.7	Sparse MSA in local region
PF05854	0.488	0.494	19.3	12.0	One beta strand orientate in a strange direction in experimental structure
PF06844	0.346	0.347	148.6	134.0	Bad orientation of C-I-TASSER model in N/C terminal due to sparse MSA
PF07352	0.315	0.313	101.0	78.4	Flexible region in experimental structure
PF07408	0.424	0.430	22.2	18.6	Low <i>Neff</i> and sparse MSA in local region
PF08127	0.251	0.298	92.7	92.3	Flexible region in experimental structure
PF08208	0.467	0.457	54.6	52.4	Flexible region in experimental structure
PF08992	0.284	0.303	15.0	7.0	Low <i>Neff</i> and flexible region in experimental structure
PF09218	0.434	0.426	9.0	7.2	Low <i>Neff</i>
PF09642	0.268	0.332	15.3	9.4	Flexible region in experimental structure
PF11428	0.295	0.318	39.2	37.6	Bad orientation of C-I-TASSER model
PF11708	0.270	0.284	11.4	11.1	Flexible region in experimental structure
PF12134	0.442	0.411	6.1	5.8	Low <i>Neff</i>
PF18484	0.362	0.348	5.1	4.9	Low <i>Neff</i>
<b>Average</b>	0.329	0.348	51.2	41.7	

## References

1. Kandathil SM, Greener JG, & Jones DT (2019) Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins* 87(12):1092-1099.
2. Ovchinnikov S, *et al.* (2017) Protein structure determination using metagenome sequence data. *Science* 355(6322):294-298.
3. Wang Y, *et al.* (2019) Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biol* 20(1):229.
4. Pfeifer F (2012) Distribution, formation and regulation of gas vesicles. *Nat Rev Microbiol* 10(10):705-715.
5. van Keulen G, Hopwood DA, Dijkhuizen L, & Sawers RG (2005) Gas vesicles in actinomycetes: old buoys in novel habitats? *Trends Microbiol* 13(8):350-354.
6. Cheng C, Bao T, & Yang ST (2019) Engineering Clostridium for improved solvent production: recent progress and perspective. *Appl Microbiol Biotechnol* 103(14):5549-5566.
7. Bourassa DV, Kannenberg EL, Sherrier DJ, Buhr RJ, & Carlson RW (2017) The Lipopolysaccharide Lipid A Long-Chain Fatty Acid Is Important for Rhizobium leguminosarum Growth and Stress Adaptation in Free-Living and Nodule Environments. *Mol Plant Microbe Interact* 30(2):161-175.
8. Cornu JY, Huguenot D, Jezequel K, Lollier M, & Lebeau T (2017) Bioremediation of copper-contaminated soils by bacteria. *World J Microbiol Biotechnol* 33(2):26.
9. Tamindzija D, *et al.* (2019) Chromate tolerance and removal of bacterial strains isolated from uncontaminated and chromium-polluted environments. *World J Microbiol Biotechnol* 35(4):56.
10. Cheng J & Charles TC (2016) Novel polyhydroxyalkanoate copolymers produced in Pseudomonas putida by metagenomic polyhydroxyalkanoate synthases. *Appl Microbiol Biotechnol* 100(17):7611-7627.
11. Li Y, *et al.* (2021) Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins*.