Supplementary Materials

Table of content

Supporting Figures

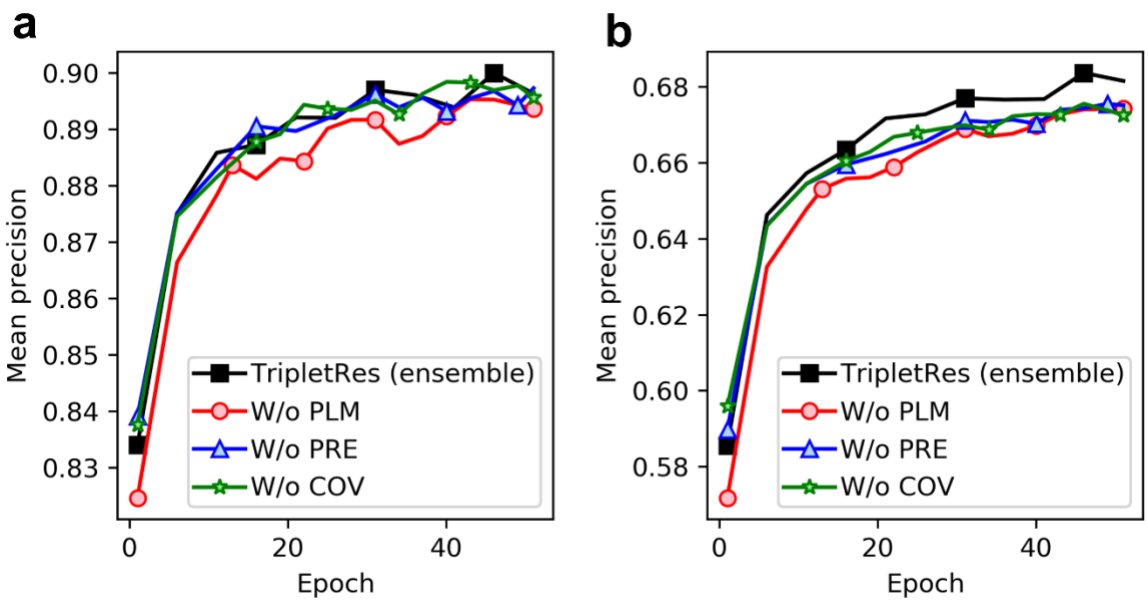

Fig S1. Comparison of the average precisions over training epochs without individual coevolutionary features and the TripletRes model that ensembles all three sets of features, on the validation set. (a) top-*L*/5 precision, (b) top-*L* precision.

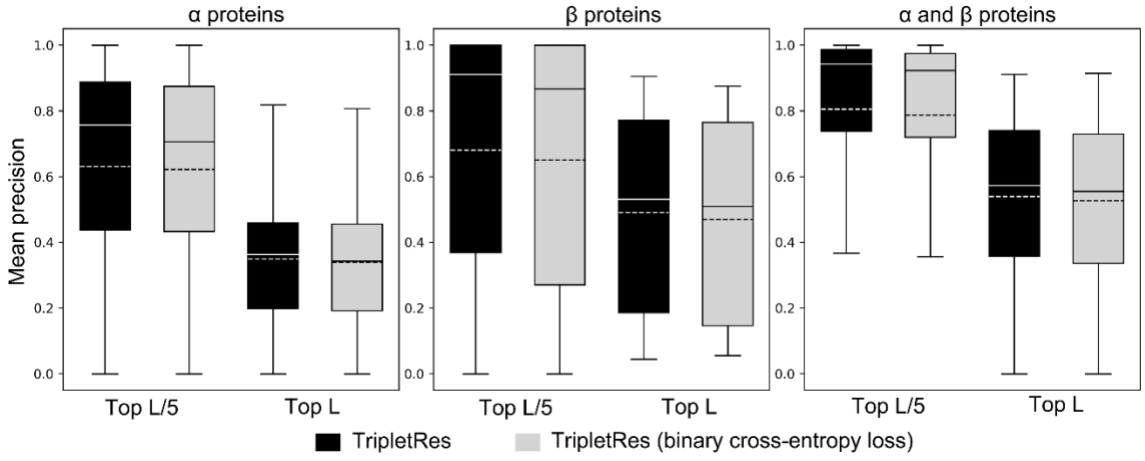

Fig S2. Comparison of long-range top-*L*/5 and top-*L* precisions with different loss functions on the different fold types, where median precision and mean precision are marked in solid and dash lines, respectively.

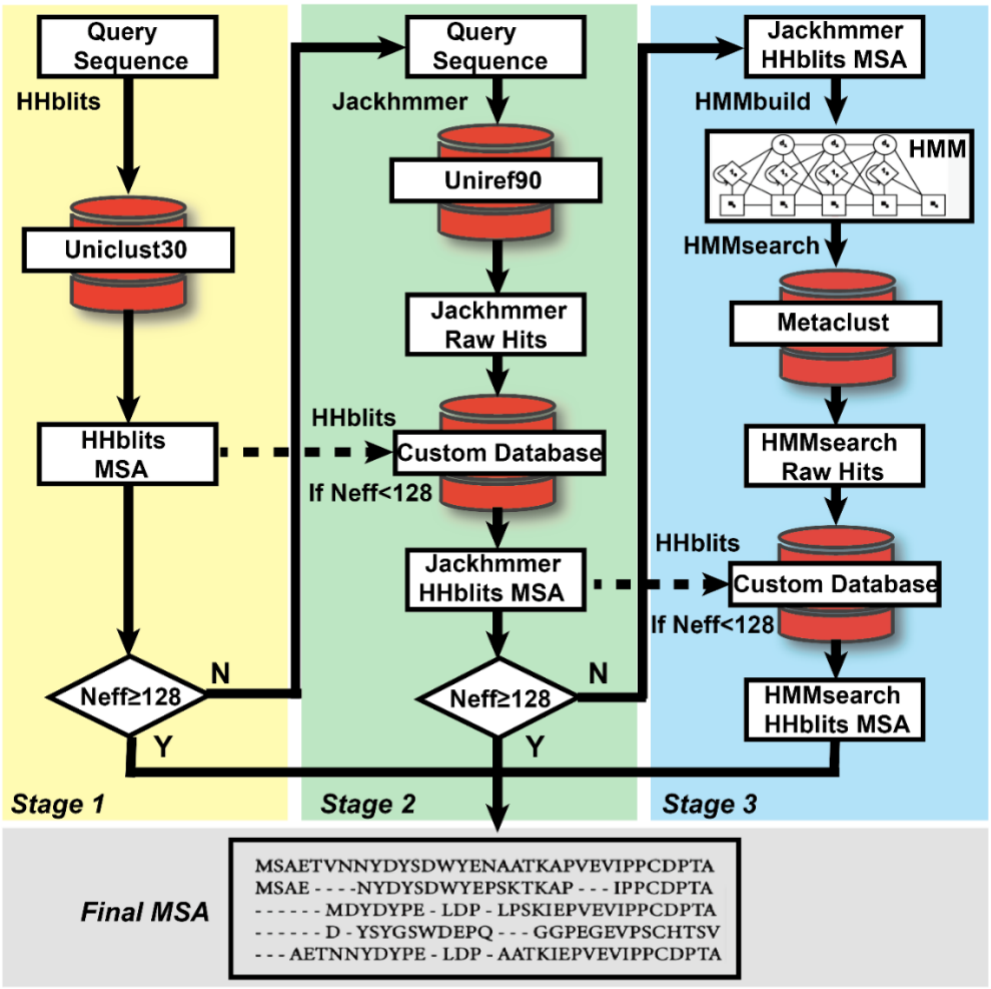

Fig S3. The DeepMSA pipeline for generating deep multiple sequence alignments for TripletRes. DeepMSA consists of three stages. The query sequence is first searched by HHblits against the Uniclust30 database to generate Stage 1 MSA (yellow background). In Stage 2 (green background), Jackhmmer searches the query sequence through the UniRef90 database to find sequence homologs, which are built into a custom database in HHblits format. HHblits is then used to search Stage 1 MSA through this custom database to get Stage 2 MSA. In Stage 3 (cyan background), the Stage 2 MSA is converted into a hidden Markov model (HMM) by HMMbuild and used by HMMsearch to search the Metaclust metagenome protein sequence database. The identified sequence homologs are reformatted to another HHblits format custom database. The MSA from Stage 2 is then used to search against this custom database to get the final stage alignment. In this incremental MSA construction process, if the MSA from either Stage 1 or Stage 2 reaches Neff>=128, this MSA will be output as final MSA without subsequent stages.

Supporting Tables

Table S1. Summary of long-range contact precision by TripletRes and control methods tweaked with DeepMSAs on 50 CASP11&12 FM targets and 195 CAMEO hard targets, sorted in ascending order of top-*L* precision. p-values in parenthesis are from a Student's t-test between TripletRes and each of the control methods, where bold fonts highlight the best performer in each category.

Methods	50 CASP FM targets				195 CAMEO hard targets			
	L/10	*L*/5	*L*/2	*L*	*L*/10	*L*/5	*L*/2	*L*
CCMpred	0.416 (1.0e-11)	0.374 (3.2e-13)	0.264 (2.6e-16)	0.187 (4.5e-17)	0.451 (1.0e-50)	0.411 (5.7e-56)	0.314 (2.8e-66)	0.229 (4.6e-67)
DNCON2	0.599 (3.5e-06)	0.551 (6.8e-06)	0.460 (2.5e-06)	0.353 (8.2e-08)	0.670 (1.4e-13)	0.622 (4.8e-18)	0.503 (2.0e-29)	0.379 (3.6e-36)
MetaPSICOV2	0.571 (2.8e-07)	0.513 (5.5e-08)	0.401 (1.9e-10)	0.299 (1.2e-12)	0.594 (7.0e-24)	0.541 (6.2e-29)	0.431 (3.4e-38)	0.323 (3.8e-42)
DeepContact	0.629 (1.3e-07)	0.583 (1.3e-06)	0.478 (1.3e-07)	0.360 (6.8e-10)	0.699 (2.8e-15)	0.643 (9.3e-21)	0.508 (2.6e-35)	0.384 (1.1e-38)
ResPRE	0.709 (2.9e-03)	0.660 (8.8e-04)	0.549 (1.6e-04)	0.429 (4.2e-05)	0.770 (6.0e-05)	0.725 (1.4e-05)	0.599 (3.8e-12)	0.457 (7.9e-18)
TripletRes	**0.771**	**0.714**	**0.597**	**0.464**	**0.801**	**0.756**	**0.637**	**0.491**

Table S2. Summary of long-range contact precision by TripletRes, TripletRes (Post-CASP13) and trRosetta based on the same MSAs on 37 hybrid test sequences.

Methods	*L*/10	*L*/5	*L*/2	*L*
TripletRes	0.775	0.698	0.578	0.434
trRosetta	**0.821**	**0.772**	**0.648**	**0.484**
TripletRes (Post-CASP13)	0.814	0.762	0.623	0.471

Supporting Texts

Text S1. Explanation that DCA models capture linear relationships between residues

Considering the pseudolikelihood maximized Potts model as an example, the marginal probability of *l*-th position in the sequence is defined by equation (5) in main text:

$$P\left(\sigma_l = \sigma_l^{(m)} \middle| \sigma_{\setminus l} = \sigma_{\setminus l}^{(m)}\right) = \frac{\exp\left(h_l\left(\sigma_l^{(m)}\right) + \sum_{k=1,k\neq l}^{L} J_{lk}\left(\sigma_l^{(m)}, \sigma_k^{(m)}\right)\right)}{\sum_{q=1}^{21} \exp\left(h_l(q) + \sum_{k=1,k\neq l}^{L} J_{lk}\left(q, \sigma_k^{(m)}\right)\right)}$$

This equation can be interpreted as a Multinomial logistic regression model, a "log-linear" model. The outcome term is the *l*-th position, and the input features are other positions except *l*-th position. $J_{lk}(a, b)$ can be considered as a regression coefficient associated with the residue type *b* at position *k* variable and the outcome (*l*-th position) with residue type *b*. $h_l(b)$ can be interpreted as the bias parameter of position *l* being residue type *b*.

In addition, the inverse of covariance matrix (precision matrix) can also be interpreted as linear regression models (1, 2).

References

1. Kwan CC. A regression-based interpretation of the inverse of the sample covariance matrix. Spreadsheets in Education. 2014;7(1):4613.
2. Stevens GV. On the inverse of the covariance matrix in portfolio analysis. The Journal of Finance. 1998;53(5):1821-7.

Text S2. Detailed procedure to collect training and test datasets

50 CASP FM targets and 195 CAMEO hard targets. The 50 non-redundant FM domains from the CASP11 and CASP12 experiments are downloaded from http://predictioncenter.org/download_area/. The 195 non-redundant targets defined as *hard* by CAMEO are collected from https://www.cameo3d.org/sp/targets/1-year/?to_date=2019-07-20 whose Submission Date is from 2018-07-28 to 2019-07-20. Both of the two test datasets have a pair-wise sequence identity < 30% within each of the two datasets. All sequences share >30% to any proteins used in training the TripletRes models are removed. For a fair comparison, those targets in test sets were also removed if they cannot be finished in 72 hours by any of the control methods.

7,671 SCOPe 2.07 domain sequences. TripletRes was trained on a subset of SCOPe 2.07 (downloaded in March 2018) domain sequences collected as per the following criteria: (1) Sequence length should be in the range of 30-400 residues; (2) Resolution of the corresponding structure should be better than 2.0 Å; (3) Maximum pairwise sequence identity is also set to 30%. There were 7,671 domains collected for training. The whole training set was split into 10 subsets, and we randomly selected one as the validation set and left the remaining subsets as the training set for hyper-parameter tuning. After the hyper-parameter tuning, the final model is the average of 10 models and each model was trained by considering each subset as the validation set and the remaining subsets as the training set.

26,151 PDB sequences. A new training set was constructed to train TripletRes, which is a non-redundant set of experimental structures from Protein Data Bank (PDB). A total number of 510,940 structures by 2019.11.12 with a maximum length of 1000 residues were initially collected. The initial sequence set was then clustered at the sequence identity threshold of 35%. The obtained 26,151 full-length sequence representatives are selected as the final training set.

37 hybrid test sequences. To objectively evaluate the performance of TripletRes trained with the new training data, a test set containing 37 sequences was constructed. The test set was the combination of 50 non-redundant FM domains from the CASP11 and CASP12 experiments, 195 non-redundant targets defined as *hard* by CAMEO and 31 FM domains in CASP13. We excluded those sequences that have a sequence identity > 40% with any sequence in the training set.

Text S3. A brief introduction of control methods and other top participants in CASP13

CCMpred (3) is a representative DCA method, and the several of other control methods, DNCON2 (4), MetaPSICOV2 (5), DeepContact (6), and ResPRE (7), are based on supervised machine learning models using outputs of CCMpred or other DCA methods as input feature. Here, DNCON2, MetaPSICOV2 and DeepContact are the top-ranking predictors in CASP12 and ranked closely to the best method, RaptorX-Contact, in CASP12 (8). ResPRE was our previous work which was built on raw precision matrix feature and shown to be comparable with many state-of-the-art methods despite the use of a single precision feature matrix. It should be noted that CCMpred does not have a built-in program for MSA generation. For a fair comparison, we tested it with the same MSAs as those used in the test phase of TripletRes. The control methods were downloaded and implemented in our local computers with default parameters and the versions of sequence databases are identical to those of DeepMSA.

In CASP13, DMP, also known as DeepMetaPSICOV (9), combines the input features of MetaPSICOV2 and a covariance feature (10) with residual convolutional neural networks (RCNNs). Meanwhile, both ZHOU-Contact, i.e. SPOT-Contact (11), and RaptorX-Contact (12) combine traditional one-dimensional features (secondary structure, solvent accessibility, and sequence profile, etc.) and pairwise coevolution features (CCMpred final output) by RCNNs or recurrent neural networks. The prediction results of other participants in CASP13 were obtained from CASP13 data archive.

References

1. Kwan CC. A regression-based interpretation of the inverse of the sample covariance matrix. Spreadsheets in Education. 2014;7(1):4613.
2. Stevens GV. On the inverse of the covariance matrix in portfolio analysis. The Journal of Finance. 1998;53(5):1821-7.
3. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014;30(21):3128-30.
4. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics. 2017;34(9):1466-72.
5. Buchan DW, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins: Structure, Function, and Bioinformatics. 2017.
6. Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing evolutionary couplings with deep convolutional neural networks. Cell systems. 2018;6(1):65-74. e3.
7. Li Y, Hu J, Zhang C, Yu DJ, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics. 2019:4647-55.
8. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins: Structure, Function, and Bioinformatics. 2018;86(S1):51-66.
9. Kosciolek T, Jones DT. Accurate contact predictions using covariation techniques and machine learning. Proteins: Structure, Function, and Bioinformatics. 2016;84(S1):145-51.

10. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics. 2018;34(19):3308-15.
11. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics. 2018;34(23):4039-45.
12. Xu J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A. 2019;116(34):16856-65.

Text S4. Traditional feature extraction strategy with post-processing

As a baseline for comparison, the traditional feature extraction method, which involves a post-processing procedure over a raw coevolutionary feature matrix, is tested. There are usually two steps in the post-processing procedure. The coevolutionary feature matrix is first transformed into an *L* by *L* contact score matrix *C* by

$$C_{ij} = \sqrt{\sum_{a,b} \left\| F_{ij}^{ab} \right\|_2^2} \tag{S1}$$

where each entry represents the potential of forming a contact. Here, *a* and *b* represent two types of amino acids, and *F* represents the obtained coevolutionary feature matrix. The contact score matrix *C* will be further normalized by an average product correction (APC) step:

$$C_{ij}^{APC} = C_{ij} - \frac{C_i C_j}{C} \tag{S2}$$

where $C_i = \frac{1}{L}\sum_{j \neq i}^{L} C_{ij}$, and $\mathrm{C} = \frac{1}{L^2 - L}\sum_{i,j,i<j}^{L} C_{ij}$. C^{APC} is the predicted contact-map based on coevolution analysis with the post-processing procedure. C^{APC} can be considered as the input feature of a supervised machine learning model. In this work, we use the same neural network structure with 22 residual blocks as the supervised learning model for the comparison of the two extraction strategies.

Text S5. Binary cross entropy loss function for training TripletRes in CASP13.

The loss function is defined as the sum of cross entropy over all the residue pairs of the training proteins:

$$\mathcal{L}_{bin} = -\sum_{t=1}^{T} y_t \log(p_t) + (1 - y_t) \log(1 - p_t) \quad \text{(S3)}$$

Here, *T* is the total number of residue pairs in the training set. $y_t = 1$ if the distance of *t*-th residue pair of native structure is below 8Å; otherwise $y_t = 0$. p_t is the predicted probability that the *t*-th residue pair forms a contact.