DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins

Chengxin Zhang, Wei Zheng, S. M. Mortuza, Yang Li, and Yang Zhang

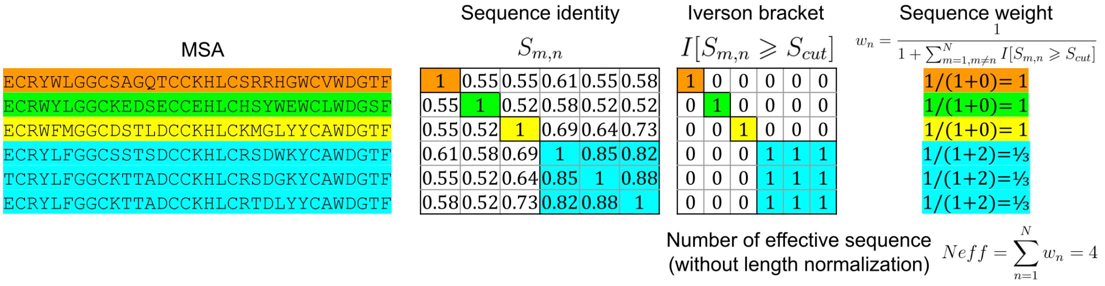
Supporting Information

Table of Content

Supporting Figures

Supporting Tables

Supporting Figures



MSA	Sequence identity $S_{m,n}$						Iverson bracket $I[S_{m,n} \geqslant S_{cut}]$						Sequence weight $w_n = \frac{1}{1+\sum_{m=1,m\neq n}^{N} I[S_{m,n} \geqslant S_{cut}]}$
ECRYWLGGCSAGQTCCKHLCSRRHGWCVWDGTF	1	0.55	0.55	0.61	0.55	0.58	1	0	0	0	0	0	1/(1+0)= 1
ECRWYLGGCKEDSECCEHLCHSYWEWCLWDGSF	0.55	1	0.52	0.58	0.52	0.52	0	1	0	0	0	0	1/(1+0)= 1
ECRWFMGGCDSTLDCCKHLCKMGLYYCAWDGTF	0.55	0.52	1	0.69	0.64	0.73	0	0	1	0	0	0	1/(1+0)= 1
ECRYLFGGCSSTSDCCKHLCRSDWKYCAWDGTF	0.61	0.58	0.69	1	0.85	0.82	0	0	0	1	1	1	1/(1+2)=⅓
TCRYLFGGCKTTADCCKHLCRSDGKYCAWDGTF	0.55	0.52	0.64	0.85	1	0.88	0	0	0	1	1	1	1/(1+2)=⅓
ECRYLFGGCKTTADCCKHLCRTDLYYCAWDGTF	0.58	0.52	0.73	0.82	0.88	1	0	0	0	1	1	1	1/(1+2)=⅓

Number of effective sequence (without length normalization) $Neff = \sum_{n=1}^{N} w_n = 4$

Figure S1. Graphic illustration for the calculation of sequence weights and the number of effective sequences. The MSA used in this example consists of $N = 6$ sequences with length $L = 33$. Using a sequence identity cutoff $S_{cut} = 0.8$, the first three sequences form three independent sequence clusters while the last three sequences form a single cluster. The four clusters are indicated by blocks colored in orange, green, yellow, and cyan in the sequence identity matrix. The Iverson bracket operation $I\left[S_{m,n} \geq S_{cut}\right]$ determines whether the sequence pair m and n has a sequence identity above the sequence identity cutoff. In other words, this operation determines whether sequence m and n are neighbors within the same cluster. We can then assign a weight for each sequence, so that the w_n weight for sequence n is inverse proportional to its number of sequence neighbor:

$$w_n = \frac{1}{1 + \sum_{m=1,m\neq n}^{N} I\left[S_{m,n} \geq S_{cut}\right]} \qquad (S1)$$

We note that any sequence n is always the sequence neighbor of itself, which hence results in the addition of one in denominator of Equation S1. The number of effective sequences (without length normalization) is:

$$Neff = \sum_{n=1}^{N} w_n \qquad (S2)$$

which equals to (1+1+1+1/3+1/3+1/3)=4 in this case. Therefore, the normalized number of effective sequences expressed in Equation 1 in the main text can be alternatively written as:

$$Nf = \frac{1}{\sqrt{L}} \cdot Neff \qquad (S3)$$

which is $1/\sqrt{33} \times 4 = 0.70$ in this case.

While our approach to calculate the number of effective sequences and sequence weights is the same as what are commonly used in many other contact prediction programs such as CCMpred, MetaPSICOV2 and TripletRes, there are also software (such as “plmc” module of the EVcoupling package) that calculates that the sequence weight and the number of effective sequences by first performing a sequence clustering. The weight of sequence n is $w_n = 1/k_n$ where k_n is the number of sequences in the sequence cluster to which sequence n belongs. This approach is equivalent to our approach because both approaches count essentially the number sequence clusters at a given sequence identity cutoff; but our approach can save the computation time needed to perform an explicit sequence clustering.

In our study, the depth of alignment was mainly quantified by the number of effective sequences instead of just the number of sequences. This is because the MSA of a query usually consists of evolutionarily related sequences sharing significant sequence similarity, and such sequence redundancy is not reflected by the number of sequences. For example, for the same query protein, the MSA with only the query protein sequence has essentially the same information as another MSA with 10 identical sequences, as both MSAs have only 1 “effective” sequence. Yet, the sequence number of the latter MSA is 10 times larger than the former.

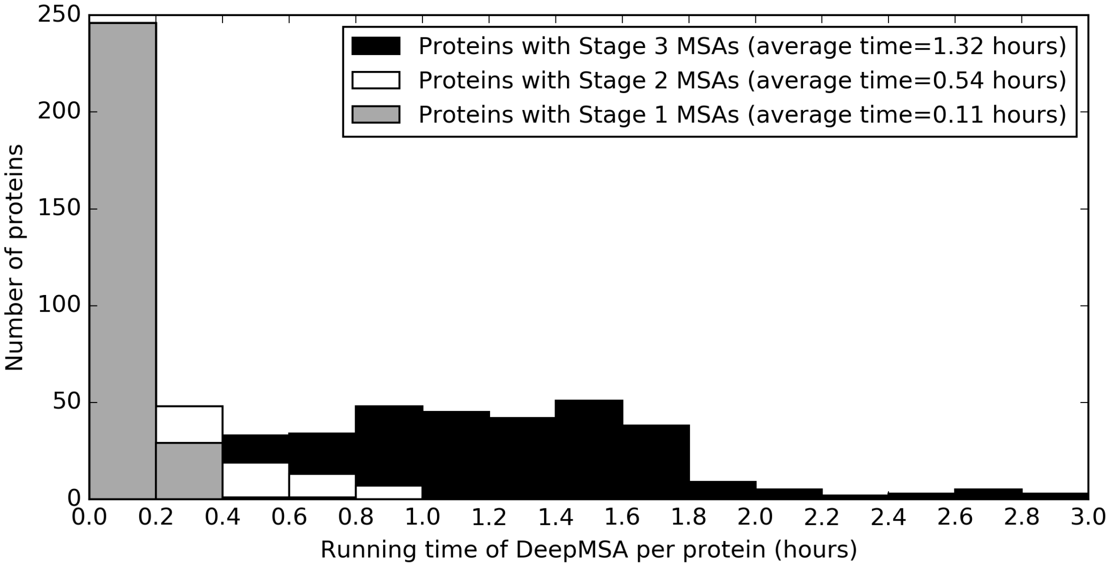


Figure S2. Stacked histogram for per protein running time of DeepMSA, with an average running time of 0.70 hour. DeepMSA does not always run all three stages to generate the final MSA. Grey, white, and black regions correspond to proteins with only Stage 1 MSA, with both Stage 1 and Stage 2 MSAs, and with Stage 1 to 3 MSAs, respectively.

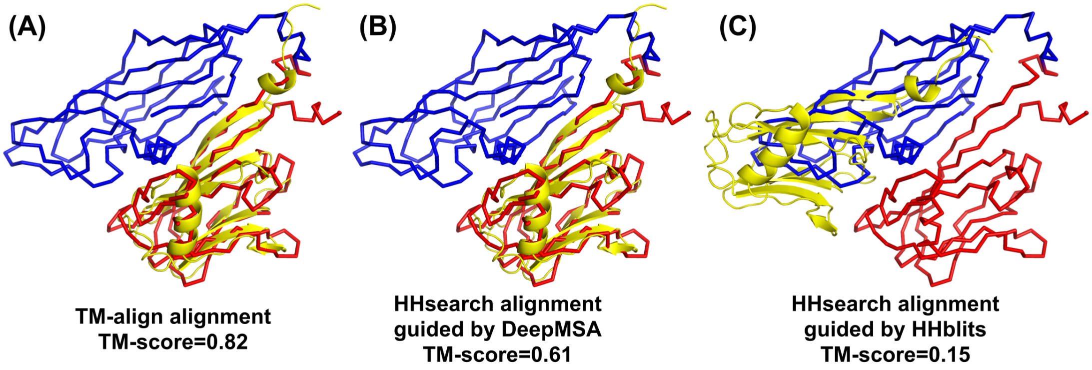


Figure S3. Alignment of query d1hx6a2 (yellow cartoon) to template 2bbdA (ribbon). The query is aligned to C-terminal of template (red ribbon at lower right) for both TM-align structure alignment **(A)** and HHsearch threading with DeepMSA profile **(B)**. However, the query is aligned to N-terminal of template (blue ribbon at upper left) for HHsearch threading using default profile **(C)**. We note that even though the query aligns to the same C-terminal region by both TM-align and DeepMSA-guided HHsearch, the sequence alignment in the latter case is slightly shifted, resulted in a lower TM-score.

Supporting Tables

Table S1. Long and medium-range contact precision for “Hard”, “Easy”, and all (“Easy” plus “Hard”) targets. Bold font indicates the higher value in each category.

Target type	Contact predictor	MSA†	Medium range contacts			Long range contacts		
			L	*L*/2	*L*/5	*L*	*L*/2	*L*/5
Hard (211)	CCMpred	DeepMSA	0.154	0.233	0.376	0.268	0.375	0.483
		Stage 1	0.136	0.199	0.303	0.215	0.307	0.410
		Stage 2	0.145	0.213	0.331	0.237	0.333	0.430
		Stage 3	**0.160**	**0.242**	**0.384**	**0.280**	**0.381**	**0.486**
		Jackhmmer	0.138	0.204	0.316	0.227	0.317	0.418
		PSI-BLAST	0.129	0.186	0.291	0.208	0.289	0.394
		No custom db	0.155	0.231	0.365	0.264	0.366	0.468
	MetaPSICOV2	DeepMSA	0.265	0.399	0.576	0.410	0.532	**0.654**
		Stage 1	0.253	0.375	0.543	0.373	0.483	0.595
		Stage 2	0.258	0.384	0.553	0.388	0.501	0.618
		Stage 3	**0.266**	**0.402**	**0.578**	**0.412**	**0.534**	0.653
		Jackhmmer	0.252	0.375	0.545	0.377	0.490	0.604
		PSI-BLAST	0.243	0.362	0.511	0.336	0.441	0.546
		No custom db	0.257	0.389	0.563	0.400	0.515	0.629
		Default	0.257	0.385	0.556	0.387	0.500	0.612
	DeepContact	DeepMSA	**0.298**	**0.451**	**0.641**	0.485	0.630	**0.756**
		Stage 1	0.285	0.426	0.600	0.445	0.581	0.716
		Stage 2	0.288	0.435	0.614	0.458	0.598	0.730
		Stage 3	0.296	0.450	0.640	**0.488**	**0.632**	0.754
		Jackhmmer	0.278	0.418	0.594	0.441	0.576	0.702
		PSI-BLAST	0.277	0.416	0.590	0.427	0.553	0.681
		No custom db	0.288	0.438	0.624	0.472	0.611	0.732
		Default	0.280	0.417	0.597	0.434	0.562	0.681
	DeepCov	DeepMSA	0.272	**0.414**	0.602	**0.439**	**0.588**	**0.738**
		Stage 1	0.262	0.395	0.572	0.408	0.553	0.701
		Stage 2	0.268	0.399	0.587	0.420	0.561	0.712
		Stage 3	**0.272**	0.411	**0.603**	0.439	0.586	0.730
		Jackhmmer	0.252	0.380	0.556	0.392	0.521	0.662
		PSI-BLAST	0.254	0.378	0.551	0.377	0.505	0.649
		No custom db	0.264	0.402	0.590	0.421	0.563	0.708
	PConsC4	DeepMSA	0.285	0.439	0.622	0.475	0.610	0.718
		Stage 1	0.269	0.403	0.574	0.420	0.544	0.653
		Stage 2	0.274	0.415	0.587	0.443	0.572	0.681
		Stage 3	**0.286**	**0.440**	**0.622**	**0.478**	**0.612**	**0.719**
		Jackhmmer	0.263	0.398	0.570	0.420	0.545	0.652
		PSI-BLAST	0.234	0.354	0.505	0.364	0.474	0.572
		No custom db	0.278	0.426	0.608	0.462	0.593	0.697
	TripletRes	DeepMSA	**0.335**	**0.525**	**0.748**	**0.610**	**0.759**	**0.860**
		Stage 1	0.330	0.516	0.734	0.594	0.742	0.849
		Stage 2	0.331	0.519	0.743	0.601	0.747	0.856
		Stage 3	0.333	0.524	0.745	**0.610**	0.756	0.859
		Jackhmmer	0.311	0.489	0.702	0.565	0.704	0.815
		PSI-BLAST	0.311	0.491	0.696	0.547	0.684	0.790
		No custom db	0.319	0.503	0.727	0.584	0.728	0.830

Easy (403)	CCMpred	DeepMSA	0.210	0.335	0.546	0.420	0.576	**0.698**
		Stage 1	0.200	0.318	0.516	0.394	0.539	0.663
		Stage 2	0.214	0.340	0.540	0.421	0.560	0.679
		Stage 3	**0.231**	**0.368**	**0.571**	**0.454**	**0.593**	0.696
		Jackhmmer	0.205	0.325	0.519	0.399	0.542	0.662
		PSI-BLAST	0.179	0.280	0.460	0.360	0.495	0.624
		No custom db	0.222	0.351	0.552	0.433	0.576	0.690
	MetaPSICOV2	DeepMSA	0.327	0.518	0.736	0.579	**0.730**	**0.849**
		Stage 1	0.322	0.503	0.718	0.558	0.703	0.822
		Stage 2	0.327	0.516	0.732	0.573	0.719	0.834
		Stage 3	**0.332**	**0.528**	**0.751**	**0.591**	0.738	0.849
		Jackhmmer	0.318	0.500	0.711	0.553	0.702	0.819
		PSI-BLAST	0.298	0.465	0.671	0.490	0.634	0.761
		No custom db	0.315	0.500	0.716	0.560	0.708	0.824
		Default	0.322	0.506	0.715	0.559	0.708	0.827
	DeepContact	DeepMSA	0.344	0.548	0.779	0.622	0.784	0.892
		Stage 1	0.340	0.540	0.764	0.606	0.766	0.878
		Stage 2	0.344	0.548	0.775	0.619	0.776	0.885
		Stage 3	**0.348**	**0.557**	**0.786**	**0.633**	**0.791**	**0.893**
		Jackhmmer	0.337	0.534	0.753	0.601	0.759	0.864
		PSI-BLAST	0.325	0.511	0.733	0.579	0.734	0.850
		No custom db	0.337	0.536	0.758	0.611	0.769	0.874
		Default	0.338	0.534	0.760	0.611	0.767	0.877
	DeepCov	DeepMSA	0.321	0.502	0.726	0.560	**0.728**	**0.860**
		Stage 1	0.319	0.500	0.720	0.552	0.718	0.854
		Stage 2	0.320	0.502	0.724	0.557	0.725	0.853
		Stage 3	**0.322**	**0.506**	**0.730**	**0.560**	0.727	0.855
		Jackhmmer	0.309	0.482	0.696	0.533	0.699	0.827
		PSI-BLAST	0.305	0.473	0.682	0.515	0.678	0.812
		No custom db	0.312	0.489	0.710	0.548	0.718	0.850
	PConsC4	DeepMSA	0.348	0.557	0.792	0.640	0.796	0.897
		Stage 1	0.339	0.544	0.772	0.620	0.773	0.877
		Stage 2	0.346	0.557	0.786	0.634	0.785	0.885
		Stage 3	**0.352**	**0.567**	**0.801**	**0.652**	**0.804**	**0.899**
		Jackhmmer	0.331	0.529	0.753	0.609	0.764	0.869
		PSI-BLAST	0.290	0.462	0.662	0.519	0.658	0.764
		No custom db	0.338	0.546	0.776	0.630	0.787	0.886
	TripletRes	DeepMSA	0.370	0.611	**0.849**	**0.725**	**0.870**	**0.941**
		Stage 1	0.368	0.607	0.845	0.720	0.866	0.937
		Stage 2	0.370	0.610	0.850	0.723	0.867	0.938
		Stage 3	**0.371**	**0.612**	0.849	**0.725**	0.868	0.936
		Jackhmmer	0.360	0.589	0.824	0.700	0.845	0.920
		PSI-BLAST	0.358	0.583	0.816	0.690	0.837	0.918
		No custom db	0.363	0.599	0.832	0.717	0.863	0.935

All (614)	CCMpred	DeepMSA	0.191	0.300	0.487	**0.368**	0.507	**0.624**
		Stage 1	0.178	0.277	0.443	0.333	0.459	0.576
		Stage 2	0.191	0.296	0.469	0.358	0.482	0.593
		Stage 3	**0.207**	**0.325**	**0.507**	0.394	**0.520**	0.624
		Jackhmmer	0.182	0.283	0.449	0.340	0.464	0.578
		PSI-BLAST	0.162	0.248	0.402	0.308	0.424	0.545
		No custom db	0.199	0.310	0.488	0.375	0.504	0.614
	MetaPSICOV2	DeepMSA	0.306	0.477	0.681	0.521	0.662	**0.782**
		Stage 1	0.298	0.459	0.658	0.494	0.627	0.744
		Stage 2	0.303	0.471	0.670	0.509	0.644	0.760
		Stage 3	**0.310**	**0.485**	**0.691**	**0.530**	**0.668**	0.782
		Jackhmmer	0.295	0.457	0.654	0.492	0.629	0.745
		PSI-BLAST	0.279	0.430	0.616	0.437	0.568	0.687
		No custom db	0.295	0.461	0.664	0.505	0.642	0.757
		Default	0.300	0.464	0.660	0.500	0.637	0.753
	DeepContact	DeepMSA	0.328	0.515	0.732	0.575	0.732	0.845
		Stage 1	0.321	0.501	0.708	0.551	0.702	0.822
		Stage 2	0.325	0.509	0.719	0.564	0.715	0.831
		Stage 3	**0.330**	**0.520**	**0.736**	**0.583**	**0.737**	**0.845**
		Jackhmmer	0.317	0.494	0.698	0.546	0.696	0.808
		PSI-BLAST	0.309	0.478	0.684	0.527	0.672	0.792
		No custom db	0.320	0.502	0.712	0.564	0.715	0.825
		Default	0.318	0.494	0.704	0.550	0.696	0.810
	DeepCov	DeepMSA	0.304	0.472	0.684	0.518	**0.680**	**0.818**
		Stage 1	0.300	0.464	0.670	0.503	0.661	0.801
		Stage 2	0.302	0.467	0.677	0.510	0.668	0.804
		Stage 3	**0.305**	**0.474**	**0.686**	**0.518**	0.679	0.812
		Jackhmmer	0.290	0.447	0.648	0.485	0.638	0.771
		PSI-BLAST	0.287	0.440	0.637	0.468	0.618	0.756
		No custom db	0.296	0.459	0.669	0.504	0.665	0.801
	PConsC4	DeepMSA	0.326	0.517	0.733	0.583	0.732	0.836
		Stage 1	0.315	0.496	0.704	0.552	0.694	0.800
		Stage 2	0.321	0.508	0.717	0.568	0.712	0.815
		Stage 3	**0.329**	**0.523**	**0.740**	**0.592**	**0.738**	**0.837**
		Jackhmmer	0.308	0.484	0.690	0.544	0.689	0.794
		PSI-BLAST	0.271	0.425	0.608	0.466	0.595	0.698
		No custom db	0.317	0.505	0.718	0.572	0.720	0.821
	TripletRes	DeepMSA	**0.358**	0.581	**0.814**	**0.686**	**0.832**	**0.911**
		Stage 1	0.355	0.576	0.807	0.676	0.823	0.907
		Stage 2	0.356	0.578	0.813	0.681	0.826	0.910
		Stage 3	0.358	**0.581**	0.813	0.686	0.829	0.909
		Jackhmmer	0.343	0.555	0.782	0.654	0.796	0.884
		PSI-BLAST	0.342	0.551	0.775	0.641	0.784	0.874
		No custom db	0.348	0.566	0.796	0.671	0.817	0.899

[†] Stage 1, 2 and 3 are three stages of DeepMSA. "No custom db" modifies DeepMSA pipeline by directly concatenating HMMER alignments without custom HHblits database construction in Stage 2 and 3. "PSI-BLAST" and "Jackhmmer" search UniRef90 with PSI-BLAST and Jackhmmer, respectively.

[*] Each *p*-value is calculated by one-tailed paired t-test to test whether DeepMSA has significant better contact prediction accuracy than the respective profile.

Table S2. Benchmark results for the first threading template for "Hard", "Easy", and all targets. Bold font indicates the higher value in each category.

Target type	Method	TM-score	*P*-value	RMSD	Coverage	#(TM-score>0.5)
Hard (211)	HHsearch	0.308	5.70E-03	11.15	0.665	33
	HHsearch$^{(D)}$	**0.331**	*	11.17	0.697	**46**
	MUSTER	0.311	7.40E-04	13.62	0.872	25
	MUSTER$^{(D)}$	**0.345**	*	12.87	0.851	**41**
Easy (403)	HHsearch	**0.691**	0.25	4.81	0.906	**370**
	HHsearch$^{(D)}$	0.689	*	5.01	0.906	369
	MUSTER	0.680	4.00E-09	5.05	0.904	358
	MUSTER$^{(D)}$	**0.687**	*	4.86	0.904	**367**
All (614)	HHsearch	0.559	1.70E-02	6.99	0.823	403
	HHsearch$^{(D)}$	**0.566**	*	7.13	0.834	**415**
	MUSTER	0.553	7.30E-10	7.99	0.893	383
	MUSTER$^{(D)}$	**0.569**	*	7.61	0.885	**408**

$^{(D)}$ indicates threading with DeepMSA profile. We note that while threading with DeepMSA profile is significantly better than that using default profile for almost all metrics, the difference is smaller for the "Easy" targets. This is partly because the default sequence profiles are deep enough for most of the "Easy" targets and the differences made from MSAs are therefore less pronounced. In fact, the results of HHsearch using DeepMSA become slightly worse than that of default HHsearch, which may be due to the parameterization of the programs that was based on the default HMM profiles.

* Each *p*-value is calculated by one-tailed paired t-test to test whether DeepMSA has significant better average TM-score than the default profile for threading.

Table S3. Benchmark results of secondary structure (SS) prediction by PSIPRED for 211 "Hard" targets. Bold font indicates the higher value in each category.

MSA	Q3	*P*-value	SOV	*P*-value
PSI-BLAST + UniRef90	82.796	1.61E-03	79.401	2.00E-03
DeepMSA	**83.616**	*	**80.601**	*

* Each *p*-value is calculated by one-tailed paired t-test to test whether DeepMSA has significant better SS prediction accuracy than the respective profile.