Published online 7 May 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W5–W10***
doi: 10.1093/nar/gkz342

MAFFT-DASH: integrated protein sequence and structural alignment

John Rozewicki[1,2], Songling Li[1,2], Karlou Mar Amada[2], Daron M. Standley[1,2,*] and Kazutaka Katoh[1,2,*]

[1]Department of Genome Informatics, Genome Information Research Center, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita 565-0871, Japan and [2]Systems Immunology Laboratory, Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita 565-0871, Japan

Received February 12, 2019; Revised April 07, 2019; Editorial Decision April 18, 2019; Accepted April 25, 2019

ABSTRACT

Here, we describe a web server that integrates structural alignments with the MAFFT multiple sequence alignment (MSA) tool. For this purpose, we have prepared a web-based Database of Aligned Structural Homologs (DASH), which provides structural alignments at the domain and chain levels for all proteins in the Protein Data Bank (PDB), and can be queried interactively or by a simple REST-like API. MAFFT-DASH integration can be invoked with a single flag on either the web (https://mafft.cbrc.jp/alignment/server/) or command-line versions of MAFFT. In our benchmarks using 878 cases from the BAliBase, HomFam, OXFam, Mattbench and SISYPHUS datasets, MAFFT-DASH showed 10–20% improvement over standard MAFFT for MSA problems with weak similarity, in terms of Sum-of-Pairs (SP), a measure of how well a program succeeds at aligning input sequences in comparison to a reference alignment. When MAFFT alignments were supplemented with homologous sequences, further improvement was observed. Potential applications of DASH beyond MSA enrichment include functional annotation through detection of remote homology and assembly of template libraries for homology modeling.

INTRODUCTION

Multiple sequence alignments (MSAs) form the basis of a wide range of biological data analyses. MSAs describe the relationships between a set of protein or nucleotide sequences that are assumed to descend from a common ancestor and thus play an integral role in our understanding of molecular evolution. MSAs also play an important role in protein structural and functional analysis. For example, detecting co-evolution from MSAs is a critical step in the prediction of protein-protein interactions (1,2) and such methods have been utilized in detection of host-pathogen interactions (3). More recently, integration of deep learning and co-evolution analysis have markedly enhanced the sensitivity of protein tertiary structure prediction (4). In the high-throughput sequencing era, scalable and accurate sequence alignment is becoming more important, but also more challenging (5).

An established approach for improving protein MSA accuracy, which was first introduced in 3DCoffee (6), is to incorporate tertiary structural information. Protein structure tends to be conserved over long evolutionary timescales even where there is no detectable homology at the sequence level (7). MSA software such as Expresso (8,9) in the T-Coffee package and PROMALS3D (10) allow structural information to be incorporated in order to improve accuracy when aligning remote sequence homologs. Since version 7, MAFFT has supported the use of structural restraints (11). Structural information can be systematically extracted from pairwise structural alignments, and this information improves alignment accuracy in benchmarks (12). Despite its contribution to alignment accuracy, however, integration of structural restraints can complicate alignment calculations due to the fact that tertiary structures are inherently higher-dimensional objects than sequences and thus core methods for their processing and alignment more elaborate. Furthermore, sequence-structure integration can often introduce additional parameters that complicate workflows and increase computational resource requirements or data storage requirements for end-users. Due to these considerations and others, tertiary-structure-restrained MSAs are far from the mainstream. For example, the vast majority of MAFFT web server queries to date have not utilized structural restraints. Thus, in order to facilitate practical use of structural information in MSAs, a number of technical challenges must be addressed.

*To whom correspondence should be addressed. Tel: +81 6 6879 8367; Fax: +81 6 6879 8368; Email: katoh@ifrec.osaka-u.ac.jp
Correspondence may also be addressed to Daron M. Standley. Email: standley@ifrec.osaka-u.ac.jp
Present address: Karlou Mar Amada, TILI.IO PTY LTD, L7 380 Docklands Drive, Docklands, Victoria 3008, Australia.

Here we have developed a web service, MAFFT-DASH (https://mafft.cbrc.jp/alignment/server/), which integrates multiple sequence alignment with a web-based database, DASH (https://sysimm.org/dash/), that serves comprehensive pairwise structural alignment information in a responsive and ready-to-use form. By employing in-house tools for structural alignment and their organization in a database of our design, we were able to circumvent the hierarchical structure imposed by CATH and SCOP, manage the flow of data from the PDB to the final result and create a maintainable up-to-date public resource. We demonstrate the utility of this approach by assessing the performance on a number of established MSA benchmark datasets.

There are a number of benefits to the MAFFT-DASH integration. In our benchmarks, MAFFT-DASH showed 10–20% improvement for MSAs of remote homologs, as measured by SP score, over standard MAFFT, and this improvement further increased when utilizing an additional option for including sequence homolog information. Importantly, there are few additional steps required for the user: the MAFFT-DASH interaction can be invoked with a single checkbox on the MAFFT web server or by a single argument (`--dash`) on the command line. In addition, because DASH alignments are pre-computed, the additional computational cost is due primarily to network overhead and mapping DASH structural alignments to MAFFT input sequences. The burden to the end user is dramatically less than that of methods that require on-the-fly structural comparisons or require the user to download and maintain a local database of structural comparisons. Taken together, in comparison with other tested software, MAFFT-DASH offers a highly convenient and efficient way of integrating sequence and structural alignment information resulting in accurate alignments with modest additional human or computational costs.

DASH DESIGN AND IMPLEMENTATION

DASH is a stand-alone web-based database of pairwise structural alignments of representative PDB entries using the RASH structural alignment method (13). DASH describes structural similarity at the residue, domain and chain levels. The domain and residue-level similarities are used by MAFFT-DASH. Representative PDB chains were defined, using CD-HIT (14), as those sharing less than 99% sequence identity. Each representative chain was decomposed into domains using Protein Domain Parser (15) and structural alignments were computed for all unique pairs. Residue-level structural similarity was defined in terms of a Gaussian function of the distance between two Cα atoms

$$S^i_{d_0} = \mathrm{e}^{-\left(\frac{d_i}{d_0}\right)^2},$$

where i is the alignment index and d_0 is a reference distance set to 4Å. Significant domain-level similarity was defined using the RASH score (13), which is a linear combination of sequence and structure-based terms that were optimized to agree with CATH and SCOP domain assignments. Full-length chain-level alignments were constructed for pairs of chains containing more than one significantly similar domain pair. This involved constructing a full length chain-chain similarity matrix composed of the residue-level structural similarities, $S^i_{d_0}$, and the BLOSUM62 amino acid exchange matrix. The sequence similarity term was used in order to generate smooth alignments in domain linker regions without residue-level structural similarity scores. The chain level alignments were computed using Needleman-Wunsch-Gotoh algorithm (16,17) on the full-length matrix. In this way, domain-domain alignments were treated rigidly, but their relative orientations via domain linker regions were treated flexibly. This was done so that the lack of structural comparison information in domain linker regions would not create artifacts or interfere with the ultimate goal of multiple sequence alignment.

DASH alignments are made available to the public in a human-readable form on the DASH website (https://sysimm.org/dash/; Figure 1D), where pairwise alignments and structures are graphically displayed in MSAViewer (18) and Molmil (19), respectively. DASH alignments are also available in a machine-readable form via a REST-like API. DASH can be searched by PDB ID, DASH Domain ID, or sequence. Data from the REST API can be sorted or filtered by most metadata columns for domains, chains, domain alignments or chain alignments. There are also separate REST API endpoints for batched sequence-based searches as a single query (up to 750 sequences) or retrieving batches of specific domain or chain alignments as a single query (up to 100 000 alignments). This is useful for users who wish to download all alignments for a specific group of domains or chains. FASTA-formatted sequence files are also provided for all DASH entries. Updates to the REST API in the future will be provided at new web addresses so as to maintain compatibility and not break tools that rely on it.

The initial pairwise alignment step involved billions of structural comparisons, but was able to be accomplished efficiently using Google Cloud Platform. The use of cloud-computing will allow the database to be updated smoothly over time to keep pace with the ever-increasing number of PDB entries.

MAFFT-DASH INTEGRATION

An additional option in the MAFFT web server and command-line tool has been developed which seamlessly incorporates DASH alignments as structural restraints for a set of input sequences (Figures 1A–C). Representative sequences are chosen by a BLAST (20) search of the DASH chain representatives. Hits for representatives for each sequence segment are then combined/filtered to make a master list of representative segments for the input set of sequences. Comprehensive structural alignments for these representative segments are then provided by DASH via the REST API. The DASH representatives are then merged with the original MAFFT input along with structural alignment-derived restraints as described below.

In the usual MSA process, group-to-group alignment is performed using dynamic programming (DP) at the progressive stage and the iterative refinement stage. For group-to-group alignment, a DP matrix is constructed using the profiles of the two sequence groups. When structural alignments exist for two groups, the residue-level equivalence scores are added to the corresponding elements of the DP

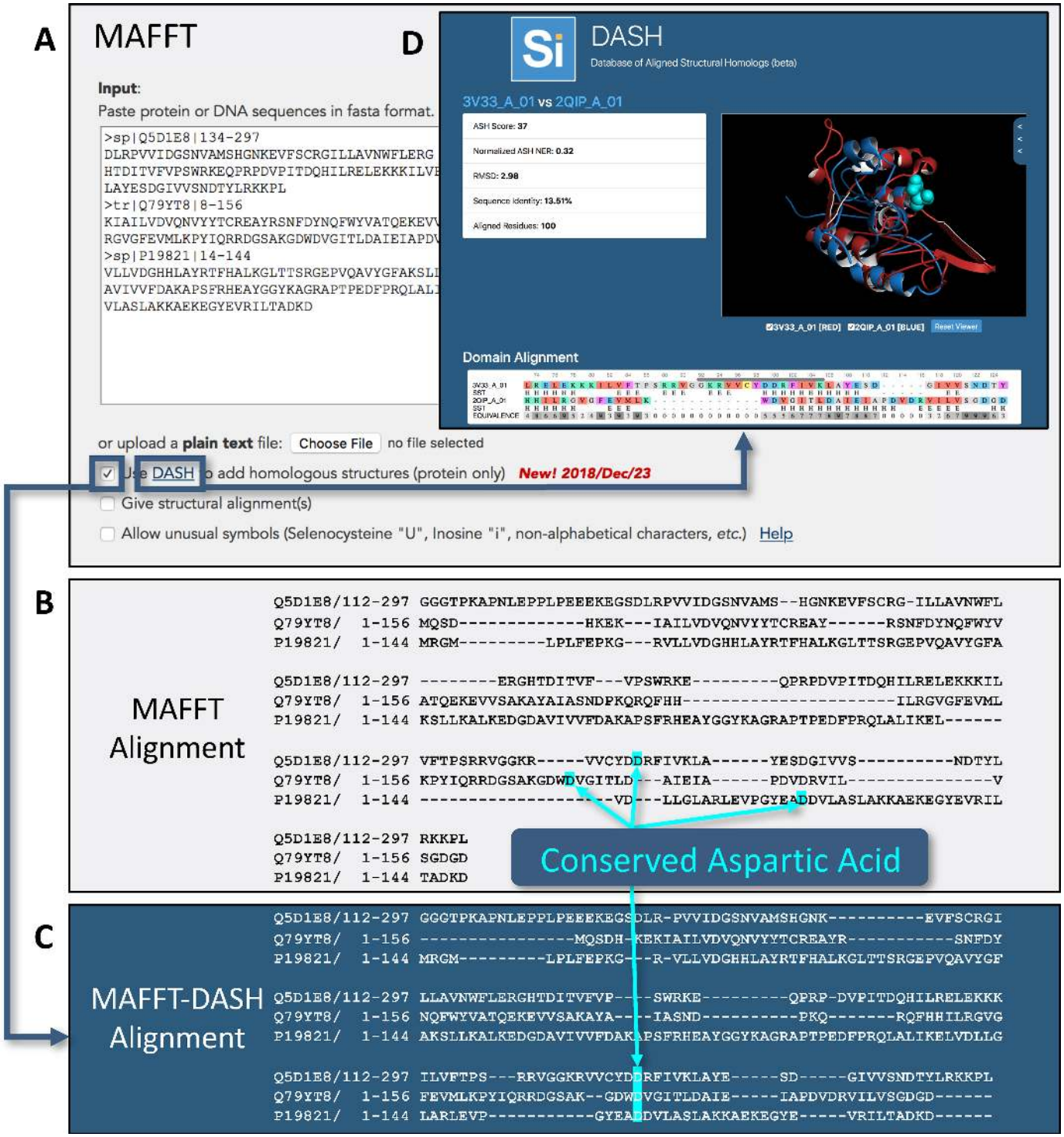

Figure 1. MAFFT and DASH Web servers. (**A**) Input consisting of three distantly-related sequences—human Regnase-1, VPA0982 from Vibrio parahaemolyticus and the nuclease domain of taq polymerase from Thermus aquaticus—with the DASH option checked. (**B**) In the default MAFFT alignment, a conserved catalytic aspartic acid is not aligned in any of the three inputs. (**C**) In the MAFFT-DASH alignment, the conserved catalytic aspartic acid is properly aligned. (**D**) The pairwise structural alignment between Regnase-1 (3V33) and VPA0982 (2QIP) as displayed in DASH.

Table 1. Benchmarks using reference MSAs

Methods \ Data	HMFM	MBSF	MBTL	OXFM	BB11	BB12	BB20	BB30	BB40	BB50	SY
						SP					
MAFFT	0.916**	0.571**	0.203**	0.894**	0.649**	0.937**	0.927**	**0.862**	**0.917**	0.899*	0.751**
Promals	0.947**	0.726**	0.475**	0.947**	**0.791**	**0.936**	0.933*	**0.883**	**0.898**	**0.903**	0.848**
T-Coffee	0.922**	0.585**	0.224**	0.909**	0.657**	**0.945**	0.916**	0.837**	**0.897**	0.895*	0.778**
Expresso	0.950**	0.708**	0.330**	0.954**	0.734**	0.903**	0.878**	0.827**	0.867**	0.874**	0.805**
MAFFT-DASH	**0.971**	0.770**	0.436**	**0.974**	0.764*	**0.943**	**0.937**	**0.880**	**0.909**	**0.918**	0.838*
MAFFT-DASH Homologs	**0.976**	**0.787**	0.530*	**0.975**	**0.793**	**0.946**	**0.938**	**0.885**	**0.889**	**0.919**	**0.851**
Promals3D	0.965**	0.780**	**0.598**	0.972**	**0.807**	0.897**	0.926**	**0.881**	**0.899**	0.899*	**0.873**
T-Coffee DASH†	0.966**	0.740**	0.396**	0.970**	0.756**	0.941*	0.934*	**0.868**	**0.899**	**0.917**	0.830**
						TC					
MAFFT	0.798**	0.254**	0.075**	0.852**	0.407**	0.838*	0.456**	**0.586**	**0.598**	0.591**	0.554**
Promals	0.851**	0.393**	0.298**	0.919**	0.582*	**0.817**	0.496**	0.516**	0.508*	0.572*	0.663**
T-Coffee	0.808**	0.262**	0.098**	0.871**	0.411**	**0.855**	0.403**	0.474**	**0.550**	**0.587**	0.591**
Expresso	0.845**	0.372**	0.173**	0.919**	0.518**	0.752**	0.369**	0.391**	0.440**	0.514**	0.579**
MAFFT-DASH	**0.909**	0.440**	0.259**	**0.961**	**0.550**	**0.853**	**0.557**	**0.610**	**0.533**	0.643*	**0.666**
MAFFT-DASH Homologs	**0.922**	**0.464**	**0.335**	**0.957**	**0.588**	**0.855**	**0.576**	**0.603**	**0.490**	**0.652**	**0.684**
Promals3D	0.892**	0.451**	**0.407**	0.952**	**0.630**	0.755**	0.502**	0.580**	0.490**	0.555**	**0.690**
T-Coffee DASH†	0.896**	0.410**	0.217**	0.950**	0.526**	**0.852**	0.466**	0.533*	**0.519**	**0.646**	0.642**
Number of cases	87	225	34	165	38	44	41	30	49	16	149

HMFM, HomFam; MBSF, Mattbench-Superfamily; MBTL, Mattbench-Twilight; OXFM, OxFam; BB11–BB50, BAliBASE subsets 11–50; SY, SISYPHUS. Scores that are significantly worse than the best are marked with * ($P < 0.05$) and ** ($P < 0.01$) as calculated with Wilcoxon signed-rank test. Others are in bold. †See the main text. Command line options are as follows: MAFFT was run with `--localpair --maxiterate 100 --thread 4 --threadit 0`. Promals and Promals3D were run with default arguments. T-Coffee was run with `-n_core 4`. Expresso was run with `-mode expresso -blast LOCAL -pdb_db '/path/to/local/pdb' -n_core 4`. MAFFT-DASH was run with `--dash --localpair --maxiterate 100 --thread 4 --threadit 0`. MAFFT-DASH Homologs was run with `mafft-homologs.rb -l -d uniref50 -o '--dash --globalpair --maxiterate 100 --thread 4 --threadit 0'`.

Table 2. Benchmarks without reference MSAs

Methods \ Data	HMFM	MBSF	MBTL
		iRMSD	
MAFFT	1.069**	2.178**	8.362**
Promals	1.025**	1.531**	**3.141**
T-Coffee	1.058**	2.107**	6.869**
Expresso	1.004**	1.607**	5.922**
MAFFT-DASH	**0.990**	1.409**	4.141**
MAFFT-DASH Homologs	**0.962**	**1.371**	**2.918**
Promals3D	0.993**	1.398**	**2.912**
T-Coffee DASH†	0.977**	1.512**	4.196**
Ideal	0.954	1.381	2.204
		Aligned NER	
MAFFT	0.804**	0.659**	0.483**
Promals	0.813**	0.692**	**0.563**
T-Coffee	0.803**	0.647**	0.488**
Expresso	0.813**	0.679**	0.511**
MAFFT-DASH	**0.817**	**0.700**	**0.549**
MAFFT-DASH Homologs	**0.818**	**0.703**	**0.566**
Promals3D	**0.817**	**0.703**	**0.573**
T-Coffee DASH†	0.813**	0.683**	0.530*
Ideal	0.819	0.714	0.611
Number of cases	87	225	34

See the footnote of Table 1 for abbreviations and symbols.

matrix. It is difficult to know *a priori* how the sequence and structural information should be weighted. We tried several different weights and confirmed that the conclusions reported here are not sensitive to the specific weight values (data not shown). MAFFT also provides an option for incorporating sequence homologs (21) and, if invoked, the homologs can be used to further query DASH alignments. DASH alignments can also be incorporated into T-Coffee as a plugin that is similar to the MAFFT-DASH integration. Preliminary results for a prototype T-Coffee plugin are described in this paper.

MSA BENCHMARKS

878 test cases were collected from the BAliBase (22), HomFam (23), OXFam (an extended version of OXBench (24)), Mattbench (25) and SISYPHUS (26) benchmark sets. HomFam and OXFam were chosen over raw HOMSTRAD and OXBench because they contain more information about reliably aligned regions (27) that can be used for more accurate scoring of estimated alignments. Extra PFAM sequences in the HomFam and OXFam datasets were removed prior to benchmarking in order to restrict the size of benchmark cases to no more than 150 sequences and

to derive a clearer assessment of the performance of methods which use sequence homologs (27).

BAliBase provides two scoring methods, Sum of Pairs (SP) and Total Columns (TC), which were calculated with `bali_score`. SP and TC scores for all other benchmark cases were computed using FastSP (28). SP compares a query MSA with a reference MSA and returns the fraction of correctly-aligned residue pairs. TC is the fraction of correctly aligned columns in the alignment.

For benchmark sets which included structure files (Mattbench, HomFam) two additional scores were calculated, iRMSD (29) and Aligned Number of Equivalent Residues (NER) (30). iRMSD is a measurement developed to evaluate the quality of MSAs when structures are known. It is the root-mean square deviation (RMSD) of distances that are less than a cutoff value (10 Å) in two structures that are composed of equivalent residue pairs. Unlike normal RMSD, it is not reliant on any specific structural superposition, and instead uses intra-molecular distance. iRMSD was computed using T-Coffee.

Because different structural alignment methods use different similarity thresholds leading to differences in the number of aligned residues, we developed an alternative MSA quality evaluation strategy, Aligned NER. Rather than being a comparison between an estimated and reference MSA, Aligned NER measures the structural accuracy of models constructed from an MSA. A model is compared with a reference structure and given a score between 0.0 and 1.0 indicating how structurally similar it is to the reference. The result is that alignment methods that align more residue pairs can be fairly compared with those that align fewer residue pairs. To compute the Aligned NER score, for each benchmark case, the sequence with known structure that is most similar to all other sequences according to BLAST is chosen as the 'query', and the rest are used as 'templates' to build homology models. Models were built using MODELLER (31) for each query/template pair using the pairwise alignment extracted from the MSA. Aligned NER was calculated using ASH by measuring the raw NER between the native template structure and the output model, and then dividing by the length of the query sequence.

Methods tested included MAFFT, PROMALS, T-Coffee, Expresso, MAFFT-DASH, MAFFT-DASH Homologs and T-Coffee-DASH (prototype). Structure databases used by PROMALS3D, Expresso, MAFFT-DASH, T-Coffee-DASH and MAFFT-DASH Homologs were blacklisted at 100% sequence identity against benchmark input sequences using BLAST in order to simulate real-world scenarios where exactly matching representatives are not available.

For reference, we additionally calculated 'ideal' benchmark scores by applying the benchmark criteria to the reference alignments. For SP and TC, which operate in alignment space, the ideal values are always perfect (1.0; not shown). However, for iRMSD and Aligned NER, which use tertiary structural information, tested programs can exceed the ideal values, suggesting that there are limitations to the use of 'reference' alignments.

The benchmark results are organized as follows: Table 1 evaluates the mean accuracy of each method (row) for each test set (column) using SP and TC scores, which depend on a reference alignment. Table 2 evaluates each method in terms of structure using iRMSD and Aligned NER. These results indicate that, overall, MAFFT-DASH Homologs performed well by all four metrics (iRMSD, Aligned NER, SP and TC). In most benchmarks, MAFFT-DASH Homologs performed best or was not significantly different from the best. The effect of structural information on alignment accuracy was clearly observed. Methods with structural information (lower half in each table) generally outperformed purely sequence-based methods (upper half), consistent with previous studies (6,10). MAFFT-DASH also significantly outperformed its sequence-based counterpart, MAFFT, in most cases.

An exceptional case is BB40 in Table 1. BB40 contains sequences with long N- and C-terminal extensions. In the current implementation, DASH will add domains that are only locally similar and do not improve the overall MSA accuracy. Thus, BB40 represents an area for improvement that will be addressed in a future release.

The average difference in runtime between MAFFT and MAFFT-DASH for the 572 benchmark cases for which there were between 5 and 150 input sequences (average of 18 sequences per case) was 124 s, making MAFFT-DASH the fastest structure-aware method tested. Among methods which combine sequence, structure, and sequence homologs, however, MAFFT-DASH Homologs took a longer amount of time (average wall-clock time of 36 minutes) than Promals3D (12 minutes). This difference is mainly because MAFFT-DASH Homologs uses newer and more comprehensive databases of sequences which increases the computational requirements when searching for homologous sequences. Reducing the overhead of adding homologous sequences without sacrificing accuracy will also be a future goal.

Based on these results we believe we were successful at achieving our goal of implementing a high performance MSA method that enables the incorporation of tertiary structural information in a painless and efficient way.

CONCLUSIONS AND FUTURE DIRECTIONS

MAFFT and its integration with DASH close the technical gaps between protein sequence and structural comparison. By leveraging cloud computing to maintain an exhaustive structural search of all PDB data, DASH has the potential to enhance many kinds of downstream analyses. Because of the low computational resource requirements and granularity of information contained in the database, we believe DASH to be particularly well-suited for large-scale analyses such as deep learning-based residue contact or distance prediction (32,33). MAFFT-DASH has also proven useful in our hands for multi-template assembly of B or T cell receptors, which share a common framework but exhibit diverse binding properties through combinatorial assembly of complementarity-determining regions (Schritt *et al.*, submitted). Current limitations include the use of a single domain decomposition algorithm in DASH, which can be addressed by establishing consensus domain boundaries (34). MAFFT-DASH will continue to be enhanced and expanded by aggregating more metadata from other sources,

making use of such data in multiple alignments and delivering results in both user- and machine-friendly ways.

ACKNOWLEDGEMENTS

The authors thank all members of the Systems Immunology Lab and Kazunori D. Yamada, Tohoku University, for discussion.

FUNDING

Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED [18am0101108j0002]; JSPS KAKENHI [16K07464]. Funding for open access charge: KAKENHI [16K07464].
Conflict of interest statement. None declared.

REFERENCES

1. de Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
2. Wang,S., Sun,S., Li,Z., Zhang,R. and Xu,J. (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.
3. Kumar,R. and Nanduri,B. (2010) HPIDB—a unified resource for host-pathogen interactions. *BMC Bioinformatics*, **11**, S16.
4. Schaarschmidt,J., Monastyrskyy,B., Kryshtafovych,A. and Bonvin,A.M.J.J. (2018) Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins*, **86**, 51–66.
5. Muir,P., Li,S., Lou,S., Wang,D., Spakowicz,D.J., Salichos,L., Zhang,J., Weinstock,G.M., Isaacs,F., Rozowsky,J. *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.*, **17**, 53.
6. O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.
7. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
8. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
9. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.M., Taly,J.F. and Notredame,C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*, **39**, W13–W17.
10. Pei,J., Kim,B.H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
11. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
12. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**, 2455–2465.
13. Standley,D.M., Toh,H. and Nakamura,H. (2007) ASH structure alignment package: sensitivity and selectivity in domain classification. *BMC Bioinformatics*, **8**, 116.
14. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
15. Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429–430.
16. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
17. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
18. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
19. https://pdbj.org/help/molmil.
20. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
21. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
22. Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.
23. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
24. Raghava,G.P., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
25. Daniels,N.M., Kumar,A., Cowen,L.J. and Menke,M. (2012) Touring protein space with Matt. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 286–293.
26. Andreeva,A., Prlic,A., Hubbard,T.J. and Murzin,A.G. (2007) SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.*, **35**, D253–D259.
27. Yamada,K.D., Tomii,K. and Katoh,K. (2016) Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics*, **32**, 3246–3251.
28. Mirarab,S. and Warnow,T. (2011) FastSP: linear time calculation of alignment accuracy. *Bioinformatics*, **27**, 3250–3258.
29. Armougom,F., Moretti,S., Keduas,V. and Notredame,C. (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.
30. Standley,D.M., Toh,H. and Nakamura,H. (2004) Detecting local structural similarity in proteins by maximizing number of equivalent residues. *Proteins*, **57**, 381–391.
31. Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics*, **54**, 5.6.1–5.6.37.
32. Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T. F. G., Qin,C., Zidek,A., Nelson,A., Bridgland,A., Penedones,H. *et al.* (2018) De novo structure prediction with deep-learning based scoring. *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction*. p. 11.
33. Xu,J. (2018) Distance-based Protein Folding Powered by Deep Learning. arXiv doi: https://arxiv.org/abs/1811.03481, 08 November 2018, preprint: not peer reviewed.
34. Heger,A., Wilton,C.A., Sivakumar,A. and Holm,L. (2005) ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res.*, **33**, D188–D191.