nature methods

ARTICLES

https://doi.org/10.1038/s41592-022-01585-1

US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes

Chengxin Zhang[1,2,3], Morgan Shine[4], Anna Marie Pyle[3,4,5] and Yang Zhang[1,6] ✉

Structure comparison and alignment are of fundamental importance in structural biology studies. We developed the first universal platform, US-align, to uniformly align monomer and complex structures of different macromolecules—proteins, RNAs and DNAs. The pipeline is built on a uniform TM-score objective function coupled with a heuristic alignment searching algorithm. Large-scale benchmarks demonstrated consistent advantages of US-align over state-of-the-art methods in pairwise and multiple structure alignments of different molecules. Detailed analyses showed that the main advantage of US-align lies in the extensive optimization of the unified objective function powered by efficient heuristic search iterations, which substantially improve the accuracy and speed of the structural alignment process. Meanwhile, the universal protocol fusing different molecular and structural types helps facilitate the heterogeneous oligomer structure comparison and template-based protein–protein and protein–RNA/DNA docking.

Structural comparison and alignment of biomacromolecules, including protein, RNA and DNA, are of fundamental importance in structural biology studies. Apart from providing intuitive visualizations of the shape comparisons, structure alignment is needed for structure-based protein function annotation[1–3], modeling mutation effects[4], rational protein design[5,6] and protein structure classification[7]. Recent applications have also been seen in the use of templates identified by structure alignment for interdomain structural assembly[8] and template-based protein–RNA docking[9].

Different methods have been developed for comparing different types of molecules. For example, Dali[10] and TM-align[11] are typical algorithms to align protein monomer structures by maximizing both alignment accuracy and coverage (the portion of aligned residues divided by the sequence length). Similarly, RNA-align[12], RMalign[13], STAR3D[14] and ARTS[15] were designed for aligning RNA and DNA molecules, while MM-align[16] was proposed to compare multichain protein complex structures. Recently, algorithms such as mTM-align[17], Matt[18] and MUSTANG[19] were proposed for aligning several protein structures. Despite their usefulness, choosing an algorithm suitable for a specific molecular alignment task can be confusing for biological users. Meanwhile, the use of different assessment matrices for different methods makes the mutual structural comparisons of different molecule types difficult.

The most widely used structural comparison matrix is the root mean square deviation (RMSD)[20] of two molecule structures. It is, however, not suitable for structure alignment because minimizing RMSD of structurally aligned regions often results in low alignment coverage. GDT[21] and MaxSub[22] were later proposed to optimize alignment accuracy and coverage simultaneously. However, both GDT and MaxSub scores are sequence length dependent, as the average score for random structure pairs has a power-law dependence on the sequence length[23], which renders the absolute magnitude of these scores meaningless. To address these issues, TM-score was proposed as the first size-independent metric by the introduction of a length-dependent scale $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ to normalize the residue distance[23,24], that is, $\text{TM-score} = 1/L \sum_{i=1}^{L_{\text{ali}}} 1/(1 + d_i^2/d_0^2)$, where L is the length of the target structure; L_{ali} is the number of aligned residue pairs; and d_i is the distance between the Cα atoms of the ith pair of aligned residues. The TM-score was recently extended to TM-score$_{\text{RNA}}$ for nucleic acid structure comparison[12] (see Supplementary Text 1 for a complete discussion on TM-score and TM-score$_{\text{RNA}}$). The unification of scoring function provides the potential to unify the structural comparison of different molecules and molecular complexes.

In this work, we developed a Universal Structure Alignment (US-align) platform, which performs three-dimensional (3D) structure alignments for monomeric and complex protein and nucleic acid structures, built on the well-established TM-score and heuristic structural alignment algorithms. The universal strategy to address all macromolecular structure alignments makes the alignments of heterogeneous complexes (such as protein–RNA complexes) feasible. Meanwhile, the extensive optimization of a uniform scoring metric enables the algorithm to generate faster and more accurate alignments compared with the state-of-the-art methods developed for specific structural alignment tasks. The source code and the online server of US-align are freely available at https://zhanggroup.org/US-align/, which accepts both legacy Protein Data Bank (PDB) and mmCIF/PDBx formats[25] and automatically recognizes and selects the optimized algorithms for different input structure types.

Results

US-align is a versatile structural alignment program that performs four different modes of alignments, each of which can handle structures of proteins, RNAs and DNAs (Fig. 1): (1) the monomeric structure alignment mode establishes the residue-level correspondences

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [2]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT, USA. [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [4]Yale Combined Program in the Biological and Biomedical Sciences, Yale University, New Haven, CT, USA. [5]Department of Chemistry, Yale University, New Haven, CT, USA. [6]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA. ✉e-mail: zhng@umich.edu

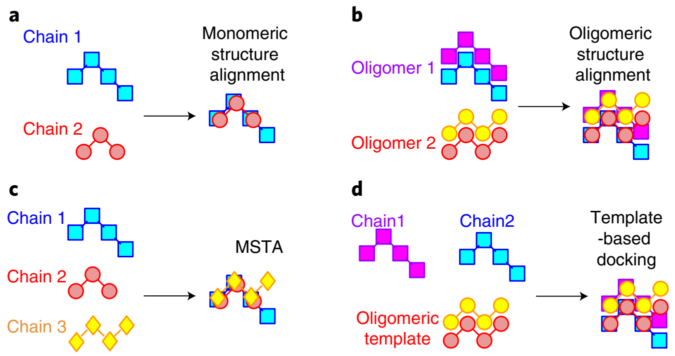


Fig. 1 | Four different structure alignment modes of US-align. a, Pairwise monomeric structure alignment. **b**, Pairwise oligomeric structure alignment. **c**, MSTA. **d**, Template-based docking of monomeric chains into an oligomeric structure. Different chains are distinguished by different colors and marker styles in this schematic.

with optimal superimposition between a pair of monomeric chains (Fig. 1a); (2) the oligomeric alignment mode establishes both chain-level and residue-level correspondences between a pair of oligomeric structures, each with two or more chains (Fig. 1b); (3) the multiple structure alignment (MSTA) mode constructs a consensus alignment from three or more monomeric structures (Fig. 1c) and (4) the template-based docking mode assembles two or more individual chains together by matching them to an oligomer template (Fig. 1d). The core idea of US-align is built on the construction of multiple heuristic alignments that cover different initial postures to avoid the trap of a specific local minimum—an issue suffered by many structural alignment methods. The follow-up rapid dynamic programming iterations help improve both accuracy and speed of the alignment procedures. The following sections benchmark the performance of US-align on the four different alignment tasks.

Oligomeric structure alignment. We first benchmarked US-align against two open-source programs for oligomeric structure alignments, MM-align[16] and MICAN[26], for oligomeric structure alignments. Whereas MM-align generates structure alignments by exhaustive combination of TM-align alignments for each individual chain pair, MICAN is built on a hierarchical strategy of secondary structure element (SSE) and residue-level alignments. The three programs were benchmarked on a set of 1,123 protein complex structures collected from the PDB that are nonredundant at a pairwise sequence identity cutoff of 30%. The dataset includes 200 dimers, 200 trimers, 200 tetramers, 129 pentamers, 200 hexamers, 60 heptamers and 134 octamers (described in detail in Supplementary Text 2).

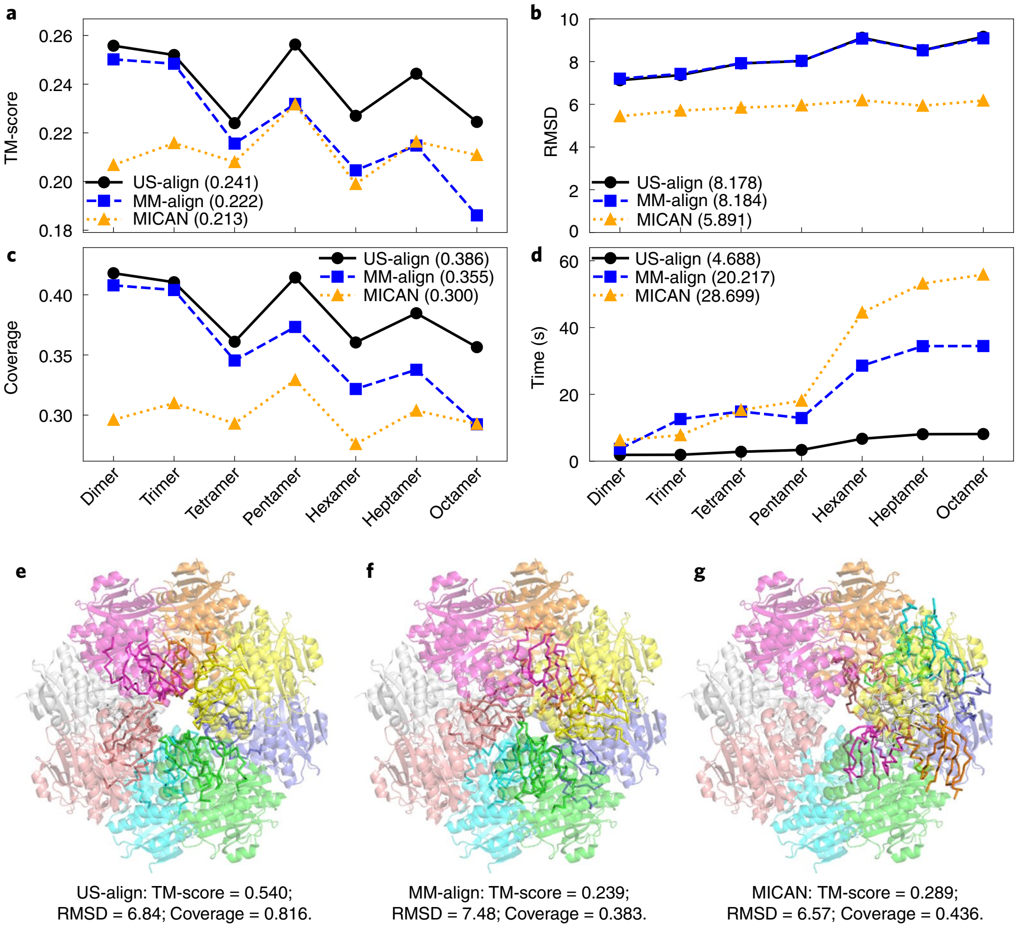
Figure 2 summarizes the performance of the three oligomeric alignment programs in terms of TM-score, RMSD, alignment coverage and execution time for all-against-all alignments among the structures with the same number of chains. As TM-score and coverage for the alignment of the same pair of structures could differ depending on whether the TM-score and coverage were normalized by the longer or the shorter structure, we reported the TM-score and coverage normalized by the shorter structure for the remainder of this manuscript, unless mentioned otherwise.

The data show that US-align consistently outperformed both MM-align and MICAN on TM-score, coverage and execution time. However, it does seem in Fig. 2b that MICAN has a lower RMSD compared with US-align and MM-align. This is because MICAN alignment covers a much smaller portion of the full structure than the other methods (Fig. 2c), which is also the reason for the low TM-score of MICAN due to the lack of balance between alignment accuracy and coverage. The average TM-score of US-align across all types of oligomers (0.243) is 8.6% and 13.1% higher than those of MM-align (0.224) and MICAN (0.215), which correspond to P values of less than 1×10^{-303} by Student's t-test.

The performance of a structural alignment method usually relies on both the alignment search engine and the objective function. The difference shown here is apparently not caused by the scoring functions, as US-align, MM-align and MICAN in this benchmark all use TM-score as the objective function. Therefore, these data highlight the efficiency of the heuristic searching process in US-align, which covers larger and more important spaces of chain assignments and structural alignments in a limited amount of CPU time. This difference was particularly evident for oligomers with more chains. For example, the average TM-score of US-align was only 2.2% higher than MM-align for the dimers but 20.6% higher than MM-align for the octamers (Supplementary Table 1). One reason for these performance differences for larger oligomers was the better ability of US-align to identify correct chain correspondences, especially for oligomers with high symmetry. If we count all 134 octamers that are the octamers of the highest number of chains in our test dataset, for example, US-align generated alignments containing, on average, 6.8 aligned chain pairs, which was 25.9% and 13.3% higher than those from MM-align (5.4) and MICAN (6.0), respectively. Figure 2e–g shows an example of the octamers from the mandelate racemase/muconate lactonizing enzyme (PDB 4JHM) and the SP_1775 protein (PDB 4IAJ) with D_4 symmetry. The optimal alignment, derived by US-align with TM-score = 0.540, aligned each of the eight chains in 4JHM to one chain in 4IAJ (Fig. 2e). On the other hand, MM-align (Fig. 2f) and MICAN (Fig. 2g) only aligned five and three out of the eight chains, respectively, leading to much lower TM-scores of 0.239 and 0.289, respectively. Although US-align generates on average more accurate oligomeric structural alignments, it could still generate suboptimal chain assignments for 2% of the cases in our test set, where one example is given in Supplementary Fig. 1 for which US-align underperforms MM-align. This is mainly because the initial chain assignment by US-align is generated by a heuristic search algorithm: the Enhanced Greedy Search (EGS; Methods). Although EGS greatly improves the speed of chain assignment with little compromise in accuracy in general, it may still very occasionally miss the best chain assignments that could otherwise be detected by an exhaustive search, such as that implemented by MM-align.

As a unique advantage of the universal structure alignment approach, US-align can perform oligomeric alignments for nucleic acid–nucleic acid or protein–nucleic acid complexes, whereas both MM-align and MICAN could deal only with protein–protein complexes. In Fig. 3, we present a case study of structure alignments between two protein–RNA complexes from two different bacteria. Since the protein components of both complexes (PDB 1Y39 chain A and PDB 2ZJR chain F) are 50S ribosomal proteins L11, they share a high structural similarity (TM-score = 0.784; Fig. 3a). Similarly, the RNA components of the two complexes (PDB 1Y39 chain C and 2ZJR chain X) are fragment and full-length 23S rRNAs, respectively, and share a high similarity (TM-score = 0.785; Fig. 3b). When combining them together, US-align creates an alignment with an even higher similarity (TM-score = 0.861; Fig. 3c) due to the cooperative optimization of the complex alignments. In Supplementary Fig. 2, we show another example where the hetero-oligomeric alignment by US-align between a protein–DNA complex and a protein–RNA complex revealed a similar mode of interaction with a significant TM-score of 0.467, which could not otherwise be captured by monomeric alignments (TM-score = 0.301 and 0.157, respectively, both below the statistical significance threshold).

Monomeric structure alignment. Structural alignments on single-chain monomer structures are a fundamental component of US-align. To examine the effectiveness of RNA monomer



US-align: TM-score = 0.540; RMSD = 6.84; Coverage = 0.816.

MM-align: TM-score = 0.239; RMSD = 7.48; Coverage = 0.383.

MICAN: TM-score = 0.289; RMSD = 6.57; Coverage = 0.436.

Fig. 2 | Performance of three oligomeric alignment programs. a–d, Performance of oligomeric protein structure alignment by US-align, MM-align and MICAN in terms of average TM-score (**a**), RMSD (**b**), alignment coverage (**c**) and running time (**d**) for complex structures in different oligomeric states (*x* axis) for $n = 200$ dimers, 200 trimers, 200 tetramers, 129 pentamers, 200 hexamers, 60 heptamers and 134 octamers. The s.e.m. values for all metrics are comparable across different methods and very small (Supplementary Table 1). Therefore, the error bars for s.e.m. are invisible. **e–g**, Octamer alignment between PDB 4JHM (semi-transparent cartoon) and PDB 4IAJ (ribbon) by US-align (**e**), MM-align (**f**) and MICAN (**g**). Each chain of the oligomer is shown in a different color.

structure comparisons, we first used CD-HIT-EST[27] to cluster the sequences of all 3,724 unique RNA chains from the PDB, resulting in 637 chains with sequence length of 30 nucleotides (nt) or more and pairwise sequence identity of less than 80%. We then ran an all-against-all pairwise alignment of these 637 chains by US-align, together with four other programs: RMalign[13], STAR3D[14], ARTS[15] and Rclick[28]. The data in Fig. 4a–d and Supplementary Table 2 show that US-align outperforms all four control RNA structure alignment programs, with a TM-score$_{\mathrm{RNA}}$ 5.8% higher than RMalign, 27.5% higher than STAR3D, 34.5% higher than ARTS and 38.6% higher than Rclick, where the difference corresponds to $P < 1 \times 10^{-303}$ for all TM-score comparisons. Furthermore, US-align is 9.6, 31.6, 2.0 and 45.7 times faster than the four control programs, respectively.

In Fig. 4e, we ran the RNA structural alignment programs to match a short rRNA-IV (PDB ID 4V8M with 135 nt) with a large 28S rRNA (PDB ID 6Y2L chain L5, 3,613 nt). Only US-align could identify the correct alignment with a TM-score = 0.595, which is 2.3 to 4.6 times higher than that identified by the other four programs. This example highlights the ability of US-align to handle RNA structure pairs with complex topologies and low sequence identities (20% in this example).

In Supplementary Fig. 3, we summarized the comparison results of US-align with four state-of-the-art monomeric protein structure alignment methods: SPalign[29], Dali[10], MICAN[26] and SSM[30]. On the 31,951 pairs of protein structures, which were collected from the all-to-all pairing of a subset of 1,000 proteins from the SCOPe database[31] 2.06, US-align creates alignments with a reasonable combination of RMSD (4.546 Å) and coverage (68.9%), resulting in the highest TM-score (0.447), which is 2.1%, 8.2%, 13.2% and 21.5% higher than that achieved by SPalign, Dali, MICAN and SSM, respectively, with $P \leq 3.4 \times 10^{-26}$ in all comparisons (Supplementary Table 3). If we count the number of the alignments with a TM-score of at least 0.5 (ref. [24]), US-align identified 8,119 pairs of similar global folds, which is 6.0%, 34.0%, 72.0% and 148.4% higher than that by SPalign (7,661), Dali (6,050), MICAN (4,720) and SSM (3,268), respectively (Supplementary Table 4). Meanwhile, the CPU time of US-align is 2.7, 6.2, 3.0 and 1.6 times lower than the benchmark programs (Supplementary Fig. 3d).

US-align also has good performance when evaluated on the objective functions from other programs, such as Q-score and Dali *Z*-score, which are unique to the SSM[30] and Dali[10] programs, respectively (see Supplementary Text 3). On the same nonredundant SCOPe dataset, US-align achieves the highest average Q-score of 0.105 (Supplementary Fig. 3e), which is 38.1%, 118.7%, 22.1% and 98.1% higher than those of SPalign, Dali, MICAN and SSM, respectively, with *P* values $\leq 5.93 \times 10^{-242}$ for all comparisons (Supplementary Table 4). Similarly, US-align achieves the second highest Dali *Z*-score (0.910), which is lower than MICAN

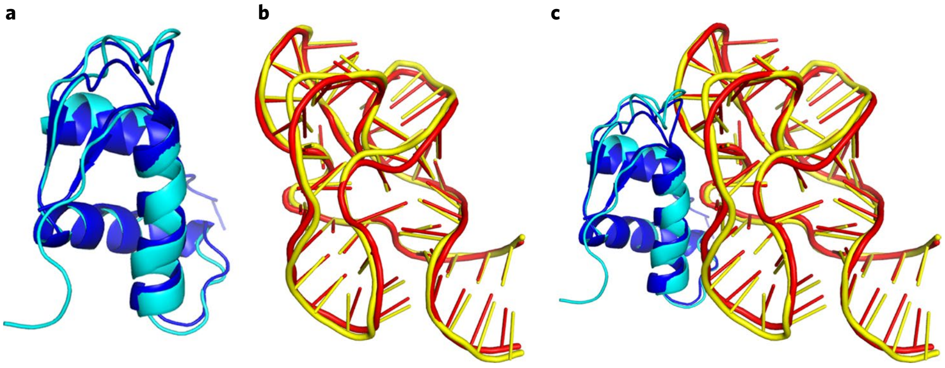


Fig. 3 | Structure alignments between two protein–RNA complexes from two different bacteria. a–c US-align alignment of protein components (**a**), RNAs (**b**) and full protein–RNA complexes (**c**) between PDB 1Y39 (blue, chain A; red, chain C) and PDB 2ZJR (cyan, chain F; yellow, chain X). Since the full 2ZJR chain X has 2,686 residues and is too large to show in the figure, **b** and **c** only show residues 1,063 to 1,119, which correspond to the region aligned to 1Y39 chain C.

(1.593) but significantly higher ($P \leq 7.48\times10^{-55}$; Supplementary Table 4) than SPalign (0.375), Dali (−1.894) and SSM (−3.175) (Supplementary Fig. 3f). The reason for the higher Dali *Z*-score by MICAN is that MICAN tends to generate alignments with lower coverage than US-align (by 19.2%). Since Dali *Z*-score is more sensitive to local variations than TM-score and Q-score, the sacrifice of alignment coverage for a smaller distance deviation at the aligned region results in a more favorable Dali *Z*-score by MICAN. Overall, the good performance of US-align on a broad range of scoring metrics reinforces the above observations that US-align has not been over-optimized for its own objective function and that its efficient alignment search engine allows it to derive reasonable alignments in a generic sense.

As an alternative assessment on the alignment performance, we calculated the agreements of manually created pairwise protein alignments from the MALIDUP dataset[32] and those from automatic protein structure alignment (Supplementary Figs. 4 and 5). Here, to avoid overfitting, we excluded MICAN, because this program was partially trained on the MALIDUP dataset[26]. The result shows that, although US-align was not optimized to resemble manual alignments, it achieves a reasonable agreement with manual alignments with an F1-score 0.782, which is 3.3%, 10.3% and 27.2% higher than those achieved by SPalign (0.757), Dali (0.709) and SSM (0.615), respectively. This probably reflects that the TM-score, which was designed to optimize the alignment accuracy and coverage simultaneously, has naturally captured the overall topological similarity of structures that is essential for the function and evolution of macromolecules.

Multiple structure alignment. MSTA matches several (three or more) monomeric structures with similar topology into a single alignment matrix. To examine the ability of US-align for RNA MSTA, we collected a benchmark dataset by clustering the 637 structures from the RNA monomer alignment dataset used above by our inhouse qTMclust algorithm (Supplementary Text 4 and Supplementary Fig. 5) at a TM-score$_{\text{RNA}}$ cutoff of 0.45. This resulted in 275, 39 and 31 clusters with one, two and more than three chains, respectively. The 31 groups with at least three structures per group were used as the MSTA benchmark dataset in which several RNA alignments were performed within each cluster.

Figure 5 shows the average performance of US-align in comparison with two third-party programs[18,19], which were extended from protein structure alignment tools (Supplementary Text 5). The comparison was based on a subset of 29 groups of RNAs for which all programs could generate results, since MUSTANG was not able to complete MSTA for two groups of long RNAs as explained in Supplementary Fig. 6. The performance on the full set of all 31 RNA groups is shown in Supplementary Fig. 6 and Supplementary Table 6. US-align outperformed the two MSTA programs (Matt and MUSTANG) by achieving 4.8% and 3.5% higher TM-score$_{\text{RNA}}$ as well as 15.5% and 63.9% lower RMSD, respectively. Here, TM-score$_{\text{RNA}}$, RMSD and coverage were all calculated from the pairwise alignments extracted from the MTSA. The result also shows that US-align was much faster than the control programs, with average times 15.0- and 1,650.3-fold shorter than Matt and MUSTANG, respectively.

As a case study, Fig. 5e shows the MTSA for a group of three RNAs: a pri-miRNA (PDB 6V5B chain D), a pre-mRNA (PDB 2L3J chain B) and a double-stranded (ds)RNA being processed by RNase III (PDB 2NUE chain C). Although all three structures have a simple topology (a single helix), MUSTANG failed to derive the correct correspondence between nucleotides of different structures, resulting in poor RMSD > 13 Å. Both US-align and Matt created correct alignments with RMSD ~3 Å, but US-align aligns more nucleotides and results in a higher coverage and TM-score$_{\text{RNA}}$ (Fig. 4e).

In Supplementary Fig. 7, we further test the ability of US-align on protein MTSA in control with four state-of-the-art methods: PROMALS3D[33], Matt[18], MAMMOTH-mult[34] and MUSTANG[19]. The benchmark dataset consists of 803 protein structures from 92 SCOPe fold families, where each fold family contains 3 to 42 structures that share the same fold but from different superfamilies[31]. Among the methods, US-align achieves the lowest pairwise RMSD (3.9 Å) with the second highest alignment coverage (68.7%). The average TM-score is 36.9–43.3% higher than other control methods, where the TM-score difference is statistically significant, with $P < 1.3\times10^{-12}$ for all comparisons (Supplementary Table 7). Meanwhile, the speed of US-align is, on average, 199.6, 24.6, 1.1 and 30.7 times faster than PROMALS3D, Matt, MAMMOTH-mult and MUSTANG, respectively.

RNA–protein docking. Given the ability of US-align for both protein and nucleic acid structure alignments, we constructed a template-based RNA–protein docking pipeline by separately matching the query RNA and protein chains to a library of known RNA–protein complex structures, with the final models sorted by the root mean square of TM-scores of the RNA and protein structural alignments.

In Fig. 6a–c, we present a summary of performance of US-align on a set of 439 nonredundant RNA–protein complexes, in comparison with two state-of-the-art RNA–protein docking methods 3dRPC[35] and PRIME[9], which perform template-free and

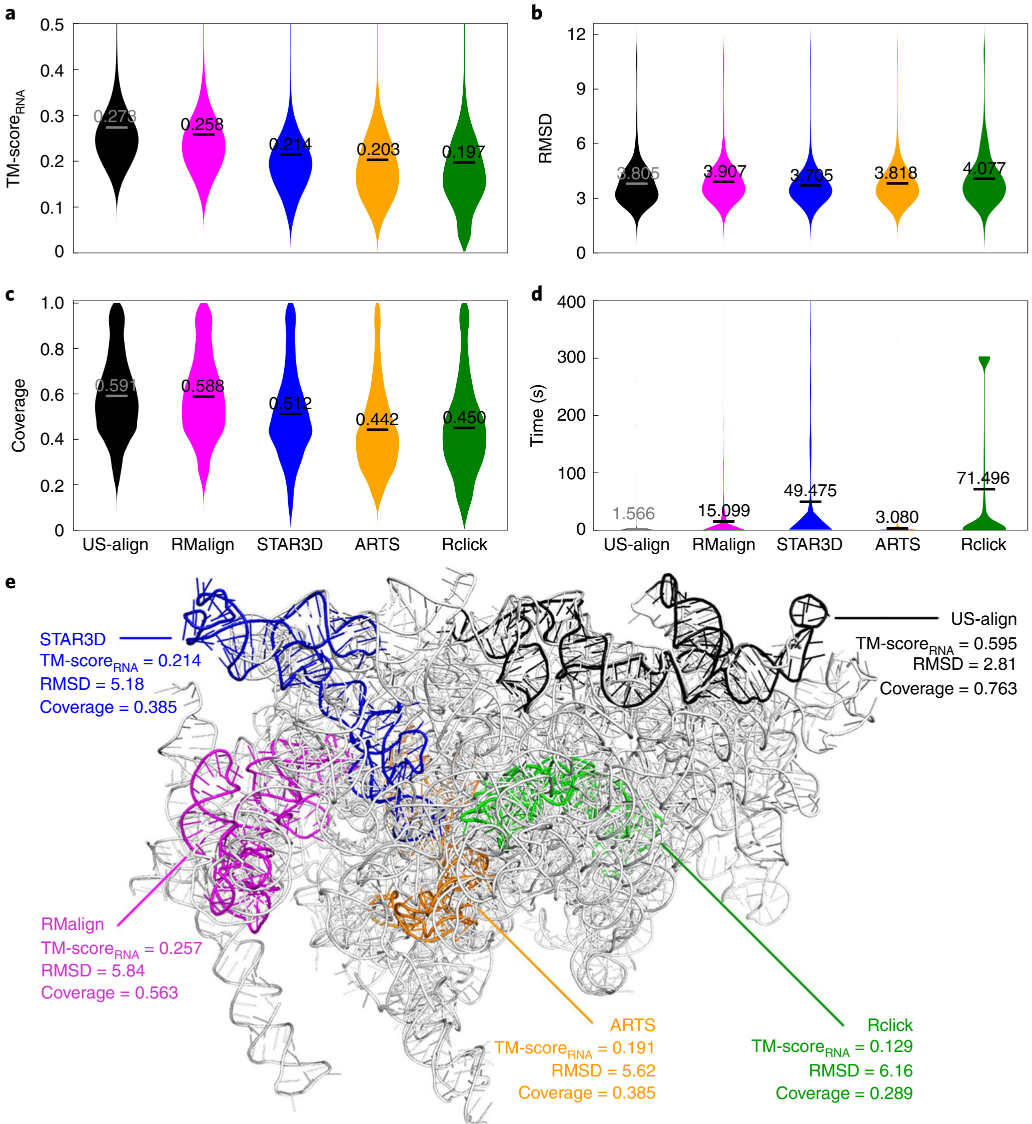


Fig. 4 | US-align outperforms four control RNA structure alignment programs. a–d, Overall performance of pairwise structure alignment of monomeric RNA chains by US-align and third-party programs, including RMalign, STAR3D, ARTS and Rclick. Since STAR3D, ARTS and Rclick generated results for only 86.8%, 84.0% and 99.8% of all pairs, respectively, this figure is for $n = 168{,}917$ chain pairs for which all programs have alignment results. The performance was measured by TM-score$_{RNA}$ (**a**), RMSD (**b**), coverage (**c**) and running time (**d**). Error bars on top of most bars were not visible because the s.e.m. values for all metrics were close to zero (Supplementary Table 2). **e**, Monomeric RNA structure alignments generated by different methods between a pair of eukaryotic rRNAs from the large ribosomal subunit: short rRNA-IV (PDB ID 4V8M chain BH, 135 nt; colored based on alignment programs) and 28S rRNA (PDB ID 6Y2L chain L5, 3,613 nt; white), with a low pairwise sequence identity of 20%.

template-based docking, respectively (Supplementary Text 6). It was shown that US-align achieved a much lower median RNA RMSD than 3dRPC (by 15.5%) and PRIME (by 22.8%). If we define a successful case as one with RNA RMSD < 10 Å, the success rate of US-align is 45.6% higher than 3dRPC and 13.8% higher than PRIME. Importantly, the average running time of US-align (19.89 min) is 28 times faster than 3dRPC (559.86 min) and 6 times faster than PRIME (118.49 min). In Fig. 6d, we present an illustrative example from the complex between a ribosomal protein and an mRNA (PDB ID 2VPL), where US-align created a model with a notably lower RMSD (1.0 Å) than 3dRPC (29.3 Å) or PRIME (8.9 Å). Although PRIME and US-align recognized the same template (PDB ID 1MZP), the US-align model is much closer to the native structure due to more precise RNA and protein structure alignments.

Discussion and conclusion

We developed US-align, a universal protocol for monomeric and oligomeric structural alignment of protein, RNA and DNA molecules, built on the coupling of a uniform TM-score objective function and the heuristic iterative searching algorithm. Large-scale benchmarks show that US-align outperforms state-of-the-art programs in terms of both alignment accuracy and speed for a wide range of structural comparison tasks, including oligomeric structural alignment, RNA and protein MSTA, and template-based protein–RNA docking. Given the fundamental importance of structure comparisons in molecular biology, the high efficiency of a uniform structural alignment tool should greatly facilitate the related structural biology and function annotation studies across different types of biomolecules.

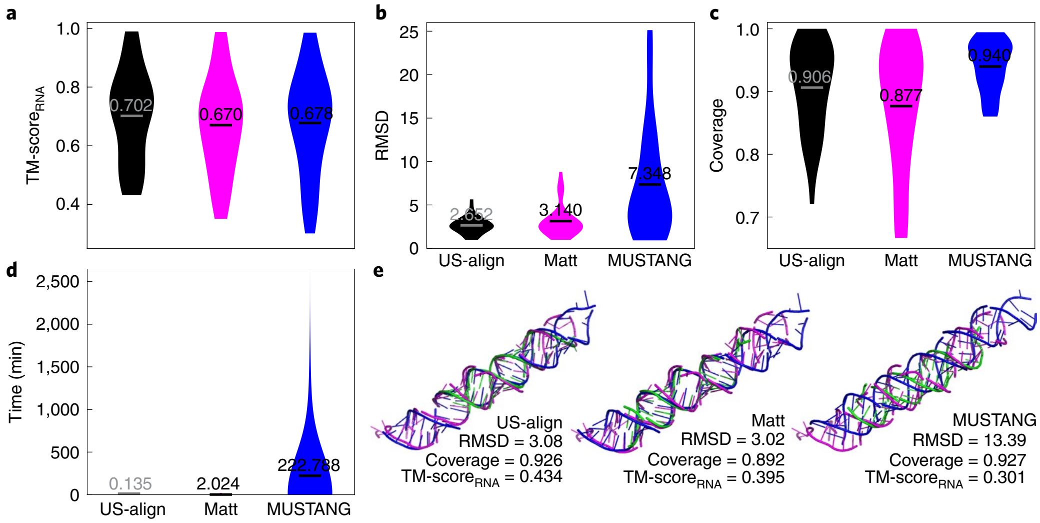


Fig. 5 | MSTA RNA alignment by US-align, Matt and MUSTANG. a–d, Average performance measured by TM-score$_{RNA}$ (**a**), RMSD (**b**), alignment coverage (**c**) and running time (**d**) for $n=29$ groups of RNAs. Horizontal lines mark the average values. The length of an error bar represents the s.e.m. Detailed per target TM-score$_{RNA}$, RMSD and coverage information is shown in Supplementary Fig. 6. **e**, Superimposed structures derived from MSTA among three RNAs: PDB 6V5B chain D (78 nt, magenta), PDB 2L3J chain B (71 nt, blue) and PDB 2NUE chain C (46 nt, green).

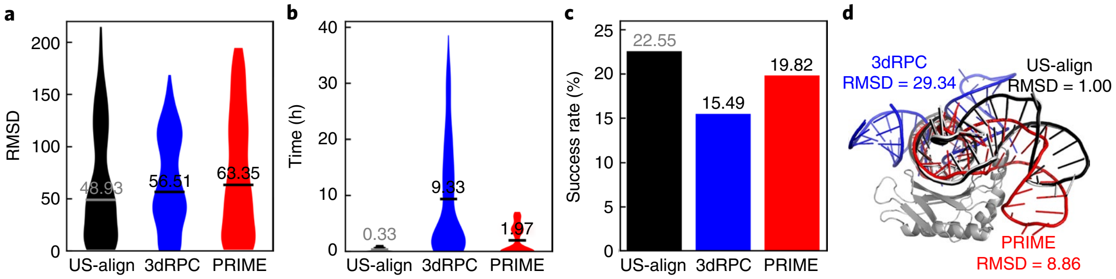


Fig. 6 | Application of US-align to RNA-protein docking. a–**c** Overall performance of RNA–protein docking by US-align (black), 3dRPC (blue) and PRIME (red) in terms of median RMSD of docked RNA ligand (horizontal line) (**a**), average running time (horizontal line) (**b**) and success rate measured by the percentage of targets with a ligand RMSD < 10 Å (**c**) for n = 439 RNA–protein complexes. Horizontal lines mark the average values. *P* values are shown in Supplementary Table 8. **d**, Example of RNA–protein docking between ribosomal protein and mRNA (chain A and B, respectively, of PDB ID 2VPL, whose native structure is shown in gray).

Despite the efficiency, US-align is essentially a tool for sequence-order dependent rigid structural alignments, which may not be sufficient for some specific applications. For example, sequence-order independent alignment is often preferred for comparing the binding pockets of ligand–receptor interactions in virtual screening studies. Meanwhile, flexible structure alignment may be needed for aligning multidomain structures with alternative interdomain orientations or for comparing multichain complexes with large conformational changes. Future developments will focus on extension of US-align for sequence-order independent and flexible alignments.

Online content

Any methods, additional references, *Nature Research* reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01585-1.

Received: 24 January 2022; Accepted: 19 July 2022;
Published online: 29 August 2022

References

1. Pazos, F. & Sternberg, M. J. Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA* **101**, 14754–14759 (2004).
2. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017).
3. Zhang, C. X., Zheng, W., Freddolino, P. L. & Zhang, Y. MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. *J. Mol. Biol.* **430**, 2256–2265 (2018).
4. Quan, L., Lv, Q. & Zhang, Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* **32**, 2936–2946 (2016).
5. Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
6. Mitra, P. et al. An evolution-based approach to de novo protein design and case study on mycobacterium tuberculosis. *PLoS Comput. Biol.* **9**, e1003298 (2013).
7. Orengo, C. A. et al. CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
8. Zhou, X. G., Hu, J., Zhang, C. X., Zhang, G. J. & Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl Acad. Sci. USA* **116**, 15930–15938 (2019).
9. Zheng, J. F., Kundrotas, P.J., Vakser, I. A. & Liu, S. Y. Template-based modeling of protein-RNA interactions.*PLoS Comput. Biol.* **12**, e1005120 (2016).

10. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480 (1995).
11. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
12. Gong, S., Zhang, C. & Zhang, Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* **35**, 4459–4461 (2019).
13. Zheng, J., Xie, J., Hong, X. & Liu, S. RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genomics* **20**, 276 (2019).
14. Ge, P. & Zhang, S. STAR3D: a stack-based RNA 3D structural alignment tool. *Nucleic Acids Res.* **43**, e137 (2015).
15. Dror, O., Nussinov, R. & Wolfson, H. J. The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.* **34**, W412–W415 (2006).
16. Mukherjee, S. & Zhang, Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* **37**, e83 (2009).
17. Dong, R., Peng, Z., Zhang, Y. & Yang, J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **34**, 1719–1725 (2018).
18. Menke, M., Berger, B. & Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.* **4**, e10 (2008).
19. Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins* **64**, 559–574 (2006).
20. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **A 32**, 922–923 (1976).
21. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
22. Siew, N., Elofsson, A., Rychiewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**, 776–785 (2000).
23. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
24. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
25. Adams, P. D. et al. Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallogr D. Struct. Biol.* **75**, 451–454 (2019).
26. Minami, S., Sawada, K. & Chikenji, G. MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C-alpha only models, alternative alignments, and non-sequential alignments.*BMC Bioinform.* **14**, 24 (2013).
27. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
28. Nguyen, M. N., Sim, A. Y. L., Wan, Y., Madhusudhan, M. S. & Verma, C. Topology independent comparison of RNA 3D structures using the CLICK algorithm. *Nucleic Acids Res.* **45**, e5 (2017).
29. Yang, Y., Zhan, J., Zhao, H. & Zhou, Y. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins* **80**, 2080–2088 (2012).
30. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268 (2004).
31. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
32. Cheng, H., Kim, B. H. & Grishin, N. V. MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins* **70**, 1162–1166 (2008).
33. Pei, J. M., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
34. Lupyan, D., Leo-Macias, A. & Ortiz, A. R. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **21**, 3255–3263 (2005).
35. Huang, Y., Li, H. & Xiao, Y. 3dRPC: a web server for 3D RNA-protein structure prediction. *Bioinformatics* **34**, 1238–1240 (2018).

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

Methods

Monomeric structure alignment. To structurally align a pair of chains in US-align, it was necessary to derive the optimal alignment (that is, the residue-level equivalence) between the two chains that maximizes the TM-score of structural superimpositions. For this, US-align starts with five sets of initial alignments:

(1) Alignments from gapless sliding of one chain against another; the alignment with the best TM-score was selected.
(2) Alignment of the secondary structures of the two chains by Needleman-Wunsch (NW) dynamic programming[36], using a gap penalty of −1, a match score of 1, and a mismatch score of 0.
(3) Alignment based on NW dynamic programming, but the matching score is a half-half combination of secondary structure match and the residue-level TM-score calculated based on the superposition from initial alignment (1):

$$\text{TM-score}_{ij} = \frac{1}{1+\left(d_{ij}/d_0\right)^2} \quad (1)$$

where d_{ij} is the distance between ith on the first structure and jth residue on the second structure.

(4) Alignments based on the superimposition of fragments with length $L_{\min}/2$ and $L_{\min}/3$, where $L_{\min}$ is the minimal length between the two query chains. To save time, a fragment is taken only every n_{jump} residues, where $n_{\text{jump}} = \min(45, L/3)$.
(5) Alignment based on gapless sliding of all continuous fragments. For proteins, a fragment is 'continuous' if it has at least four residues and all Cα–Cα distances between adjacent residues are less than 4.25 Å. For nucleic acids, any 4 nt adjacent in the sequence are considered a piece of continuous fragments.

Each of the initial alignments is followed by a heuristic iteration alignment process, in which we first rotate the structures by TM-score rotation matrix based on the aligned residues in the initial alignment. Next, a new alignment is derived using NW dynamic programming, based on the residue-level matching score (equation 1, calculated from the new superposition) with a gap penalty of −0.6. The new alignment will result in a newer superposition that will be used to create a newer alignment. The process is repeated until convergence, where the structural alignment with the highest TM-score is returned. The overall procedure of monomeric structure alignment is illustrated by Supplementary Fig. 8.

Oligomeric structure alignment. One challenge to oligomer complex alignment is chain equivalence assignment, that is, finding the correct chain-level correspondence. In the simplest scenario of aligning two dimers, US-align needs two separate structural alignments: one for aligning chains A and B from dimer 1 to chains A and B in dimer 2, respectively; and another for aligning chains A and B from dimer 1 to chains B and A in dimer 2, respectively. More generally, when aligning an oligomer with C_1 chains to another oligomer with C_2 chains where $C_1 \geq C_2$, US-align needs to determine the best chain assignment with the highest TM-score out of all $C_1!/(C_1 - C_2)!$ possible chain assignment combinations. For example, alignment of a pair of octamers requires consideration of $8!/(8-8)! = 40,320$ possible chain assignments. This makes the exhaustive search approach, such as that used by MM-align[16], extremely time consuming.

Therefore, when aligning oligomers with three or more chains, US-align employs a light-weighted chain assignment method. First, all-against-all chain-to-chain alignments are performed between all chains in oligomer 1 and all chains in oligomer 2 using fTM-align[37]—a fast version of TM-align. Compared with the standard TM-align, fTM-align decreases the number of iterations, thereby greatly reducing the required computing time while maintaining TM-scores highly correlated with those from standard TM-align[37], especially for very large structures (Supplementary Fig. 9). The TM-scores from fTM-align are then used for initial chain assignment using the EGS[38] algorithm, by maximizing the sum of the TM-scores for all assigned chain pairs (Supplementary Fig. 10).

Once an initial chain assignment is decided, US-align will perform a TM-score superimposition of the two oligomers according to the interoligomer residue-level alignments generated in the previous step (Supplementary Fig. 11). Next, based on the TM-score superimposition matrix (equation 1), a new optimal structure alignment will be obtained by a modified NW dynamics program that ignores the regions of unassigned chains (Supplementary Fig. 12). Given the new structural alignment, the chain-to-chain TM-scores will be computed for all interchain pairs and used by EGS to determine a new set of chain assignments (Supplementary Fig. 10), which will be returned to the last step for oligomer alignment iterations. This iteration will be repeated until convergence, where the structural alignment with the highest TM-score encountered during the iteration will be returned as the final structural alignment of the input oligomers (Supplementary Fig. 11).

Multiple structure alignment. To create a uniform alignment for several structures, US-align first performs all-against-all alignments among all input structures to obtain the pairwise TM-scores. Next, a structure-based guide tree is constructed based on the TM-scores using the extended unweighted pair group method with arithmetic mean (UPGMA) algorithm[39] (Supplementary Fig. 13). Finally, the pairwise alignments from the first step are progressively merged into a single MSTA according to the branching order of the UPGMA tree. To merge an alignment with M structures to another alignment with N structures, NW dynamic programming is performed using a generalized version of the residue-level TM-score from equation (1):

$$\text{TM-score}(M, N)_{ij} = \sum_{m=1}^{M}\sum_{n=1}^{N}\frac{1}{1+\left(\frac{d_{ij}(m,n)}{d_0}\right)^2} \quad (2)$$

where $d_{ij}(m,n)$ is the distance of ith residue position of the mth structure in the first alignment and jth residue position of the nth structure from the second alignment after the superposition. The overall workflow of MSTA is illustrated by Supplementary Fig. 14.

Template-based docking. To perform template-based docking, US-align implements a subroutine to align several query chains to one complex template. In this subroutine, each query chain is aligned to every chain of the complex template to calculate the TM-score. Based on the TM-scores, each query chain is then superimposed to one of the template chains so that no more than one query chain is assigned to the same template chain and the overall docking TM-score is maximized:

$$\text{TM-score}_{\text{dock}} = \sqrt{\frac{1}{2K}\sum_{k=1}^{K}\left(\text{TM}^2_{k,\text{query}} + \text{TM}^2_{k,\text{template}}\right)} \quad (3)$$

where K is the total number of query chains; $\text{TM}_{k,\text{query}}$ and $\text{TM}_{k,\text{template}}$ are the TM-score (or $\text{TM-score}_{\text{RNA}}$) for aligning the kth query chain, as normalized by the chain length of query and template, respectively.

When several templates were available, we ran US-align template-based docking on each complex template, and the template with the highest $\text{TM-score}_{\text{dock}}$ was used to generate the final docking result.

Statistics. All P values are calculated by two-tailed paired Student's t-test implemented by SciPy v.1.2.1, NumPy v.1.16.6 and Python v.2.7.18.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data needed to reproduce this work are available at https://doi.org/10.6084/m9.figshare.16725745 under CC BY v.4.0. Source data are provided with this paper.

Code availability

An online webserver and the standalone program of US-align are available at https://zhanggroup.org/US-align. The latest source code of US-align is also available at https://github.com/pylelab/USalign, while the source code for US-align version 20220227 used by this manuscript is included in Supplementary Software. The code was tested on Linux, Windows and Mac OS, where no notable differences in speed across different operating systems were found (Supplementary Fig. 15).

References

36. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
37. Dong, R., Pan, S., Peng, Z., Zhang, Y. & Yang, J. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.* **46**, W380–W386 (2018).
38. Hu, J., Liu, Z., Yu, D. J. & Zhang, Y. LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics* **34**, 2209–2218 (2018).
39. Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**, 1409–1438 (1958).

Acknowledgements

We thank Y. Cao for technical assistance in developing qTMclust and X. Wei for insightful discussions. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI1548562. C.Z. is a Howard Hughes Medical Institute postdoctoral fellow. A.M.P. is a Howard Hughes Medical Institute Investigator. This work is supported in part by the National Human Genome Research Institute (HG011868 to A.M.P.), National Institute of General Medical Sciences (GM136422, OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.) and the National Science Foundation (IIS1901191, DBI2030790, MTM2025426 to Y.Z.).

Author contributions

Y.Z. conceived the study. C.Z., A.M.P. and Y.Z. designed the experiments. C.Z. developed the method. C.Z. and M.S. drafted the manuscript. All authors revised the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01585-1.

Correspondence and requests for materials should be addressed to Yang Zhang.

Peer review information *Nature Methods* thanks Ruth Nussinov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Arunima Singh in collaboration with the *Nature Methods* team.

nature portfolio

Corresponding author(s): Yang Zhang

Last updated by author(s): May 21, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NumPy (1.16.6) and Python (2.7.18) were used in data collection.
Data analysis	SciPy (1.2.1), NumPy (1.16.6), Python (2.7.18) and US-align (20220227 available at https://doi.org/10.6084/m9.figshare.16725745) were used in data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data needed to reproduce this work are available at https://doi.org/10.6084/m9.figshare.16725745 under CC BY 4.0.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study used 1123 protein complexes for oligomer structure alignment; 637 chains for monomeric RNA structure alignment; 1000 SCOPe domains for monomeric protein structure alignment; 439 protein-RNA complexes for protein-RNA docking; and 31 groups of RNAs for multiple structure alignment. Sample sizes are chosen to encompass all structures satisfying the selection criteria, as detailed in the manuscript.
Data exclusions	A structure is excluded if it is redundant to any other structure in the same dataset with <30% (or <80%) protein (or RNA) sequence identity. These sequence identity cutoffs are commonly used by previous studies.
Replication	All results can be reproduced by our online web-server and standalone programs on Linux, MacOS and Windows.
Randomization	None. Covariants are controlled by excluding redundant structures at sequence identity cutoffs of 30% and 80% for proteins and RNAs.
Blinding	There was no blinding group or analysis in this manuscript.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging