SUPPLEMENTARY MATERIAL

SupeRNAlign: a new tool for flexible superposition of homologous RNA structures and inference of accurate structure-based sequence alignments

Paweł Piątkowski[1], Jagoda Jabłońska[2], Adriana Żyła[1], Dorota Niedziałek[1], Dorota Matelska[1], Elżbieta Jankowska[1], Tomasz Waleń[1,3], Wayne K. Dawson[1] and Janusz M. Bujnicki[1,2,*]

[1] Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Trojdena 4, 02-109 Warsaw, Poland

[2] Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznań, Poland

[3] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

[*] To whom correspondence should be addressed. Email: iamb@genesilico.pl

PP and JJ contributed equally to this work

Supplementary Table 1: Performance of individual superposition/alignment methods (median SPS scores for each family)

Family name	Rfam ID	Struc-tures	Super-positions	SETTER	iPARTS	ARTS	SARA	LaJolla	Rclick	R3D Align	SupeRNAlign	R3D Align*	SARA-Coffee*	SA-Coffee*
5S_rRNA	RF00001	9	36	0.4479	0.6348	0.6385	0.7480	0.7341	0.7150	0.7362	0.7649	0.7510	0.7560	0.7570
AdoCbl-variant	RF01689	2	1	0.8269	0.9231	0.9423	0.9423	0.9038	0.9423	0.9423	0.9423	0.9327	0.9519	0.9327
Archaea_SRP_1	RF01857	2	1	0.4058	0.6304	0.6522	0.7174	0.7681	0.7246	0.6957	0.8623	0.9203	0.8261	0.9130
Archaea_SRP_2	RF01857	2	1	0.2313	0.9552	0.9552	0.9552	0.9627	0.9552	0.9552	0.9552	0.9552	0.9328	0.9552
Bacteria_small_SRP	RF00169	3	3	0.4216	0.5294	0.8922	0.9020	0.8922	0.8922	0.5294	0.7647	0.6373	0.9216	0.7353
c-di-GMP-I	RF01051	3	3	0.2952	0.7909	0.6857	0.7872	0.7340	0.6915	0.8381	0.8000	0.8364	0.8727	0.7660
Cobalamin	RF00174	2	1	0.2816	0.7184	0.7379	0.6553	0.7087	0.7233	0.6990	0.7379	0.7621	0.7816	0.7330
glmS	RF00234	3	3	0.2966	0.3793	0.3219	0.3724	0.3793	0.3724	0.6621	0.7103	0.6414	0.5685	0.7448
Glycine	RF00504	2	1	0.4318	0.4318	0.8466	0.1932	0.8580	0.8523	0.5000	0.2022	0.4438	0.8523	0.1818
HDV_ribozyme	RF00094	2	1	0.1382	0.2033	0.2033	0.2033	0.2033	0.2033	0.2033	0.2033	0.2358	0.1870	0.1870
Histone3	RF00032	2	1	1	1	1	1	1	1	1	1	0.9643	1	0.9286
HIV-1_DIS	RF00175	2	1	0.6087	0	-	0.6087	0.6087	0.6087	0	0	0	1	0.0435
Intron_gpI	RF00028	3	3	0.4425	0.4955	0.6725	0.7178	0.8318	0.7344	0.5087	0.4775	0.3874	0.5993	0.3844
Intron_gpII	RF00029	2	1	0.5294	0.2059	0	0.4412	0.2647	0.0588	0.3824	0.8824	0.9412	0.8235	1
IRES_HCV	RF00061	4	6	0.7307	0.7832	0.8897	0.7771	0.8416	0.7940	0.7764	0.7971	0.8639	0.6651	0.7260
LSU_rRNA		3	3	-	-	0.8541	-	-	0.8293	0.8156	0.8525	0.7881	-	0.8453
Lysine	RF00168	2	1	0.2775	0.9364	0.9827	0.9711	0.9480	0.9306	0.9595	0.9827	0.9653	0.9769	0.9653
Metazoa_SRP	RF00017	5	10	0.5618	0.6446	0.7858	0.6172	0.7444	0.6921	0.7061	0.6325	0.6741	0.5013	0.4535
pfl	RF01750	3	3	0.5075	0.1791	0.1351	0.5224	0.5672	0.6418	0.1757	0.4478	0.6119	0.5405	0.4030
PreQ1	RF00522	2	1	1	1	1	1	1	1	1	1	1	1	1
Purine	RF00167	7	21	0.6761	0.9296	0.9710	0.9718	0.9420	0.9577	0.9718	0.9718	0.9577	1	0.9718
RNaseP_bact_b	RF00011	2	1	0.6076	0.7180	0.9099	0.6337	0.9099	0.5523	0.6599	0.7122	0.6192	0.4157	0.5610
SAM	RF00162	6	15	0.2419	0.5917	0.8250	0.8258	0.8333	0.8468	0.5833	0.7167	0.7177	0.8629	0.7417
SSU_rRNA		2	1	0.2044	-	0.7153	-	-	0.7339	0.6407	0.8753	0.8946	-	0.8850
THF	RF01831	2	1	0.0891	0.3960	0.4950	0.4950	0.4950	0.4059	0.4752	0.4257	0.5347	0.4257	0.5050
tmRNA	RF00023	4	6	0.8259	0.7576	0.6036	0.8153	0.9123	0.9538	0.9472	0.9538	0.8336	0.9472	0.9297
TPP	RF00059	3	3	0.1379	0.7241	0.7356	0.7356	0.7356	0.7356	0.7356	0.7356	0.7126	0.8046	0.8161
tRNA	RF00005	57	1596	0.5063	0.6337	0.6974	0.7500	0.7375	0.7600	0.7792	0.8182	0.7733	0.8204	0.8375
tRNA-Sec	RF01852	2	1	0.6809	0.5957	0.6702	0.6915	0.6915	0.7021	0.5000	0.6383	0.4681	0.7234	0.5957
U6	RF00026	2	1	0.2222	0.2778	0.1000	0.2333	-	0.2222	0.2222	0.2778	0.2556	0.0222	0.2889
ydaO-yuaA	RF00379	4	6	0.2334	0.7626	0.8373	0.8257	0.8237	0.8335	0.8400	0.8360	0.8360	0.8074	0.7961
yybP-ykoY	RF00080	2	1	0.1638	0.5431	0.5603	0.5776	0.5862	0.5690	0.4655	0.3362	0.1379	0.5862	0.3448
Whole dataset		151	1734	0.5057	0.6400	0.7020	0.7530	0.7435	0.7631	0.7791	0.8181	0.7750	0.8203	0.8354

Asterisks (*) denote sequence alignments yielded directly by the program (not generated by pdb3aln based on the superimposed structures)

Supplementary Table 2: Performance of individual superposition/alignment methods (median 3SP scores for each family)

Family name	Rfam ID	Struc-tures	Super-positions	SETTER	iPARTS	ARTS	SARA	LaJolla	Rclick	R3D Align	SupeRNAlign	R3D Align*	SARA-Coffee*	SA-Coffee*
5S_rRNA	RF00001	9	36	0.3039	0.5655	0.6208	0.7318	0.7040	0.7107	0.6733	0.7462	0.7894	0.7071	0.7594
AdoCbl-variant	RF01689	2	1	0.7370	0.8586	0.8976	0.8976	0.8490	0.8976	0.8976	0.8976	0.9075	0.9171	0.9075
Archaea_SRP_1	RF01857	2	1	0.3424	0.5013	0.6633	0.6959	0.7445	0.6298	0.6036	0.8381	0.9020	0.7619	0.8868
Archaea_SRP_2	RF01857	2	1	0.3072	0.9670	0.9670	0.9670	0.9494	0.9670	0.9670	0.9670	0.9563	0.9345	0.9670
Bacteria_small_SRP	RF00169	3	3	0.2108	0.3599	0.8032	0.8796	0.8032	0.8361	0.4552	0.7157	0.7710	0.8179	0.7486
c-di-GMP-I	RF01051	3	3	0.2316	0.7429	0.6429	0.7770	0.7504	0.6957	0.8190	0.8000	0.8357	0.8190	0.7663
Cobalamin	RF00174	2	1	0.2053	0.6979	0.7076	0.6180	0.6769	0.6923	0.6882	0.7480	0.7520	0.7376	0.7617
glmS	RF00234	3	3	0.2607	0.4579	0.4839	0.5277	0.5217	0.4582	0.3676	0.4584	0.3573	0.4184	0.4839
Glycine	RF00504	2	1	0.2159	0.2159	0.7718	0.4148	0.7775	0.7898	0.2500	0.4119	0.4643	0.7595	0.3636
HDV_ribozyme	RF00094	2	1	0.2999	0.6016	0.6016	0.6016	0.5247	0.6016	0.6016	0.6016	0.5987	0.5935	0.5935
Histone3	RF00032	2	1	1	1	1	1	1	1	1	1	0.9821	1	0.9643
HIV-1_DIS	RF00175	2	1	0.3043	0	-	0.3043	0.3043	0.3043	0	0	0	0.5000	0.0217
Intron_gpI	RF00028	3	3	0.2494	0.2488	0.4433	0.5469	0.5879	0.5966	0.2772	0.2678	0.3959	0.5420	0.2584
Intron_gpII	RF00029	2	1	0.4076	0.2458	0.1786	0.4706	0.1681	0.1008	0.2269	0.9055	0.9706	0.6975	1
IRES_HCV	RF00061	4	6	0.3654	0.3996	0.6817	0.4698	0.4357	0.4346	0.4064	0.4642	0.5127	0.3325	0.3998
LSU_rRNA		3	3	-	-	0.8322	-	-	0.7971	0.7850	0.8410	0.7964	-	0.8265
Lysine	RF00168	2	1	0.2689	0.9203	0.9913	0.9719	0.9397	0.9311	0.9524	0.9913	0.9690	0.9884	0.9827
Metazoa_SRP	RF00017	5	10	0.3857	0.3929	0.4593	0.5234	0.4894	0.4670	0.4005	0.5960	0.5987	0.4267	0.5061
pfl	RF01750	3	3	0.5454	0.1104	0.0884	0.5557	0.5343	0.6334	0.1920	0.5364	0.7018	0.5454	0.5140
PreQ1	RF00522	2	1	1	1	1	1	1	1	0.5000	1	0.5000	1	1
Purine	RF00167	7	21	0.6343	0.9267	0.9630	0.9655	0.9365	0.9594	0.9630	0.9681	0.9514	0.9815	0.9514
RNaseP_bact_b	RF00011	2	1	0.3142	0.3590	0.6945	0.3794	0.6945	0.4116	0.3299	0.3561	0.5492	0.2391	0.2805
SAM	RF00162	6	15	0.1210	0.5181	0.8292	0.8453	0.8356	0.8375	0.4356	0.7334	0.7694	0.8525	0.7458
SSU_rRNA		2	1	0.1778	-	0.7016	-	-	0.7263	0.6131	0.8726	0.9026	-	0.8913
THF	RF01831	2	1	0.0446	0.3786	0.5947	0.5809	0.5809	0.4807	0.4876	0.5879	0.6840	0.4490	0.6691
tmRNA	RF00023	4	6	0.6970	0.6537	0.6293	0.7384	0.8208	0.9141	0.8905	0.9180	0.8008	0.8525	0.8711
TPP	RF00059	3	3	0.0690	0.8103	0.8103	0.8504	0.8398	0.8103	0.4944	0.8504	0.5050	0.8333	0.7874
tRNA	RF00005	57	1596	0.4398	0.5793	0.6694	0.7211	0.7089	0.7359	0.6626	0.8106	0.7580	0.7788	0.8429
tRNA-Sec	RF01852	2	1	0.6833	0.6122	0.7351	0.7315	0.7600	0.7653	0.5357	0.7477	0.7055	0.7760	0.7407
U6	RF00026	2	1	0.1111	0.1701	0.2063	0.3354	-	0.2049	0.1111	0.2639	0.2840	0.0111	0.2382
ydaO-yuaA	RF00379	4	6	0.2225	0.7662	0.8453	0.8236	0.8417	0.8462	0.4539	0.8285	0.4516	0.8376	0.7980
yybP-ykoY	RF00080	2	1	0.1819	0.6841	0.6802	0.6888	0.6806	0.6845	0.5703	0.5681	0.5440	0.6306	0.5849
Whole dataset		151	1734	0.4317	0.5812	0.6737	0.7246	0.7120	0.7383	0.6621	0.8078	0.7627	0.7802	0.8399

Asterisks (*) denote sequence alignments yielded directly by the program (not generated by pdb3aln based on the superimposed structures)

Supplementary Table 3: Differences of median SPS between programs.

	SETTER	iPARTS	ARTS	SARA	LaJolla	RClick	R3D Align	SupeRN Align	R3D Align*	SARA-Coffee	SupeRN Align-Coffee
SETTER	0	-0.134	-0.196	-0.247	-0.238	-0.257	-0.273	-0.312	-0.269	-0.315	-0.330
iPARTS	0.134	0	-0.062	-0.113	-0.104	-0.123	-0.139	-0.178	-0.135	-0.180	-0.195
ARTS	0.196	0.062	0	-0.051	-0.042	-0.061	-0.077	-0.116	-0.073	-0.118	-0.133
SARA	0.247	0.113	0.051	0	0.009	-0.010	-0.026	-0.065	-0.022	-0.067	-0.082
LaJolla	0.238	0.104	0.042	-0.009	0	-0.020	-0.036	-0.075	-0.031	-0.077	-0.092
RClick	0.257	0.123	0.061	0.010	0.020	0	-0.016	-0.055	-0.012	-0.057	-0.072
R3D Align	0.273	0.139	0.077	0.026	0.036	0.016	0	-0.039	0.004	-0.041	-0.056
SupeRNAlign	0.312	0.178	0.116	0.065	0.075	0.055	0.039	0	0.043	-0.002	-0.017
R3D Align*	0.269	0.135	0.073	0.022	0.031	0.012	-0.004	-0.043	0	-0.045	-0.060
SARA-Coffee	0.315	0.180	0.118	0.067	0.077	0.057	0.041	0.002	0.045	0	-0.015
SupeRNAlign-Coffee	0.330	0.195	0.133	0.082	0.092	0.072	0.056	0.017	0.060	0.015	0

Legend:

$0.001 \leq P < 0.05$

$1 \cdot 10^{-10} \leq P < 0.001$

$P < 1 \cdot 10^{-10}$

Values in the table reflect the differences of median SPS scores between programs in the left column and programs in the upper row. Colored background means that selected program from the left column scored significantly better than a program from the upper row (different colors denote levels of P-values, as described in the legend.)

Supplementary Table 4: Differences of median RMSD between programs.

	SETTER	ARTS	iPARTS	LaJolla	SARA	RClick	R3D Align	SupeRN Align
SETTER	0	-1.537	-1.899	-2.155	-2.311	-2.532	-2.554	-2.838
ARTS	1.537	0	-0.362	-0.618	-0.774	-0.995	-1.017	-1.301
iPARTS	1.899	0.362	0	-0.256	-0.412	-0.633	-0.655	-0.939
LaJolla	2.155	0.618	0.256	0	-0.156	-0.377	-0.399	-0.682
SARA	2.311	0.774	0.412	0.156	0	-0.221	-0.243	-0.526
RClick	2.532	0.995	0.633	0.377	0.221	0	-0.022	-0.306
R3D Align	2.554	1.017	0.655	0.399	0.243	0.022	0	-0.284
SupeRNAlign	2.838	1.301	0.939	0.682	0.526	0.306	0.284	0

Legend:

$0.001 \leq P < 0.05$

$1 \cdot 10^{-10} \leq P < 0.001$

$P < 1 \cdot 10^{-10}$

Values in the table reflect the differences of median RMSD scores between programs in the upper row and programs in the left column. Colored background means that selected program from the left column scored significantly better than a program from the upper row (different colors denote levels of P-values, as described in the legend.)

Supplementary Table 5: Comparison of SPS scores obtained by SupeRNAlign with and without ClaRNet.

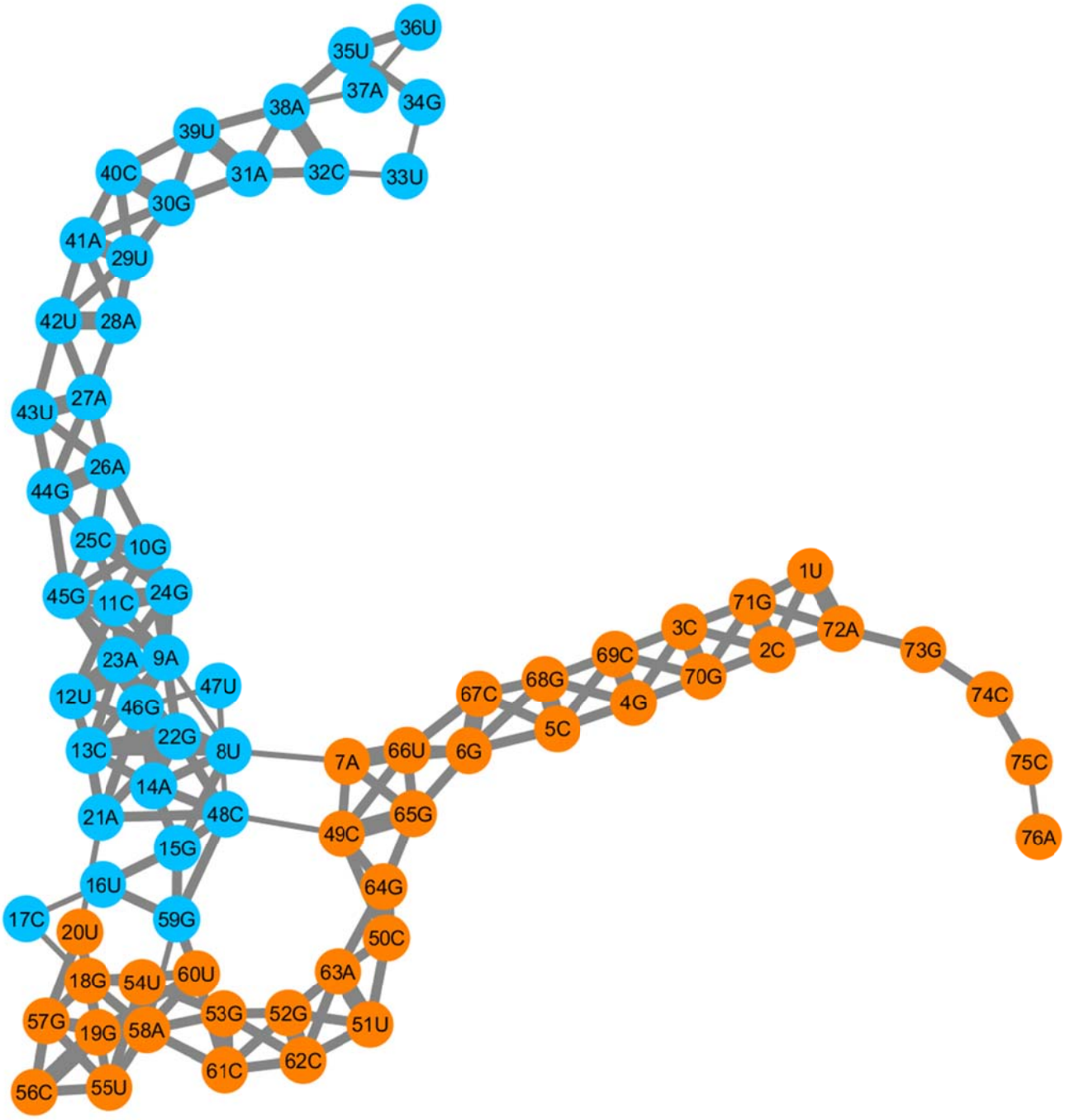
Aligner	**P-value**
SupeRNAlign-ARTS	9.854023e-10
SupeRNAlign-LaJolla	0.001078868
SupeRNAlign-R3D Align	7.541349e-11
SupeRNAlign-SARA	0.999994
SupeRNAlign-Setter	0.003803946

Values in the table show the results of Wilcoxon test for SPS scores obtained by SupeRNAlign (with different superposition tools) with and without a preliminary clustering by ClaRNet. Null hypothesis: results returned by SupeRNAlign with ClaRNet are not better than without ClaRNet.

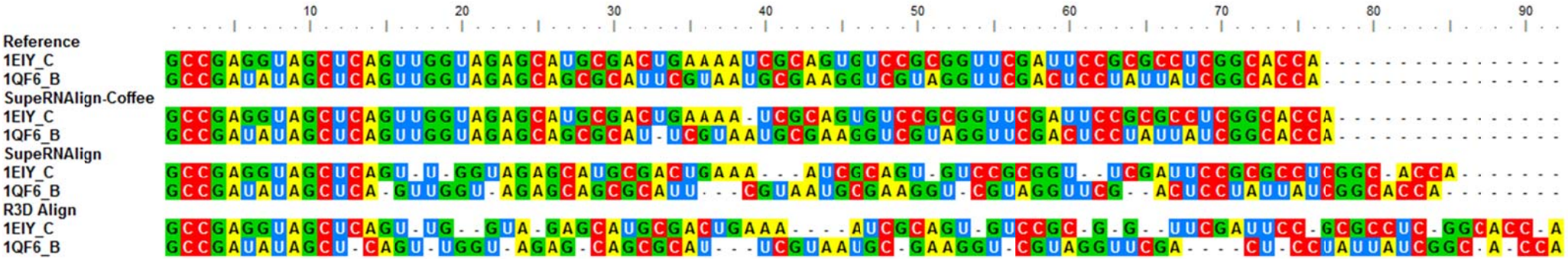
Supplementary Table 6: Parameters of ClaRNet.

Parameter	**Value**
MCL inflation parameter (*I*)	1.3
Detected contacts threshold	0.6
Contacts scores	
covalent	0.5
canonical_bp	2.0
non-canonical_bp	1.0
stacking	1.0
base-phosphate	1.0
base-ribose	1.0
other	1.0

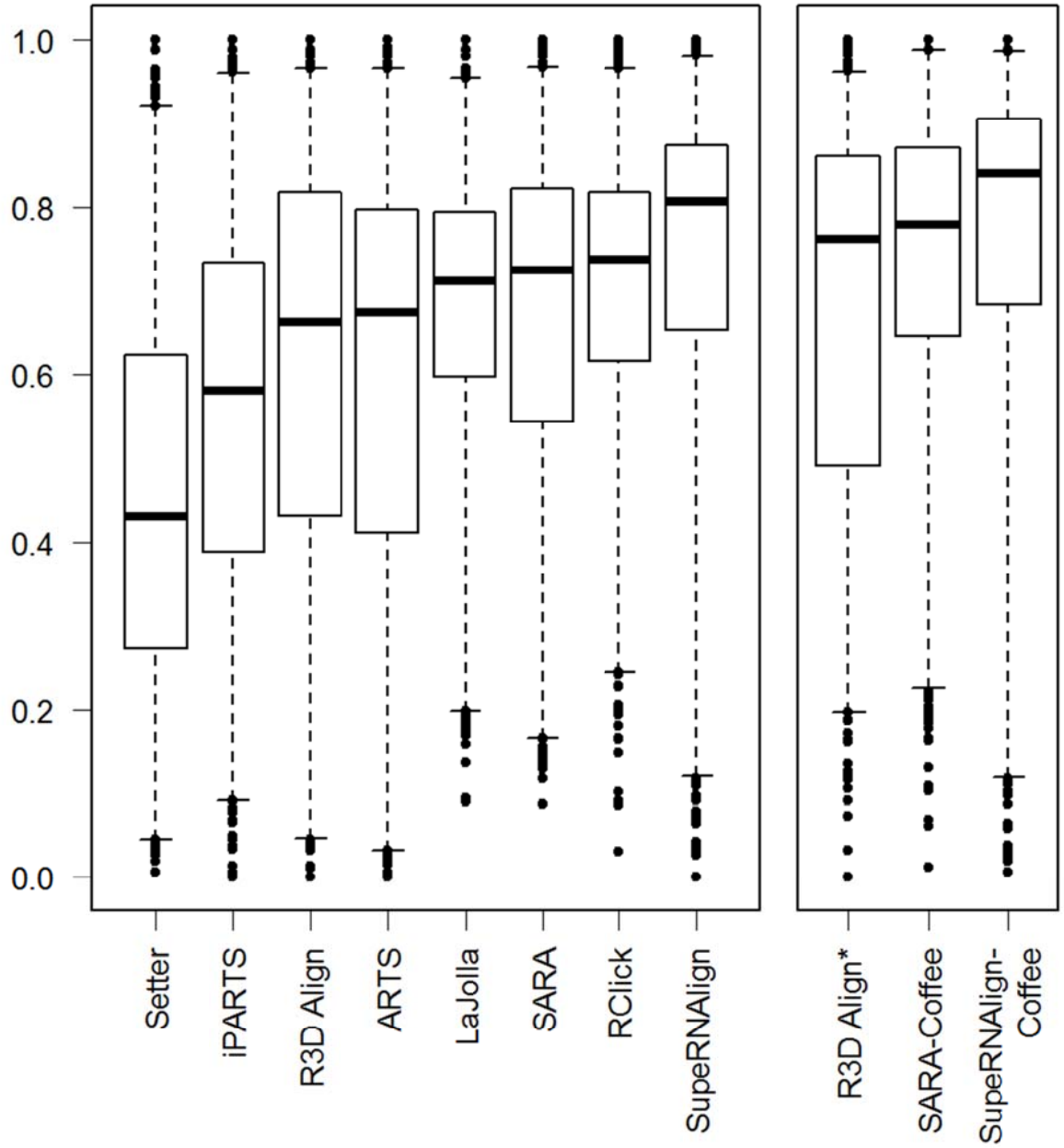
Values of default parameters used in ClaRNet procedure.



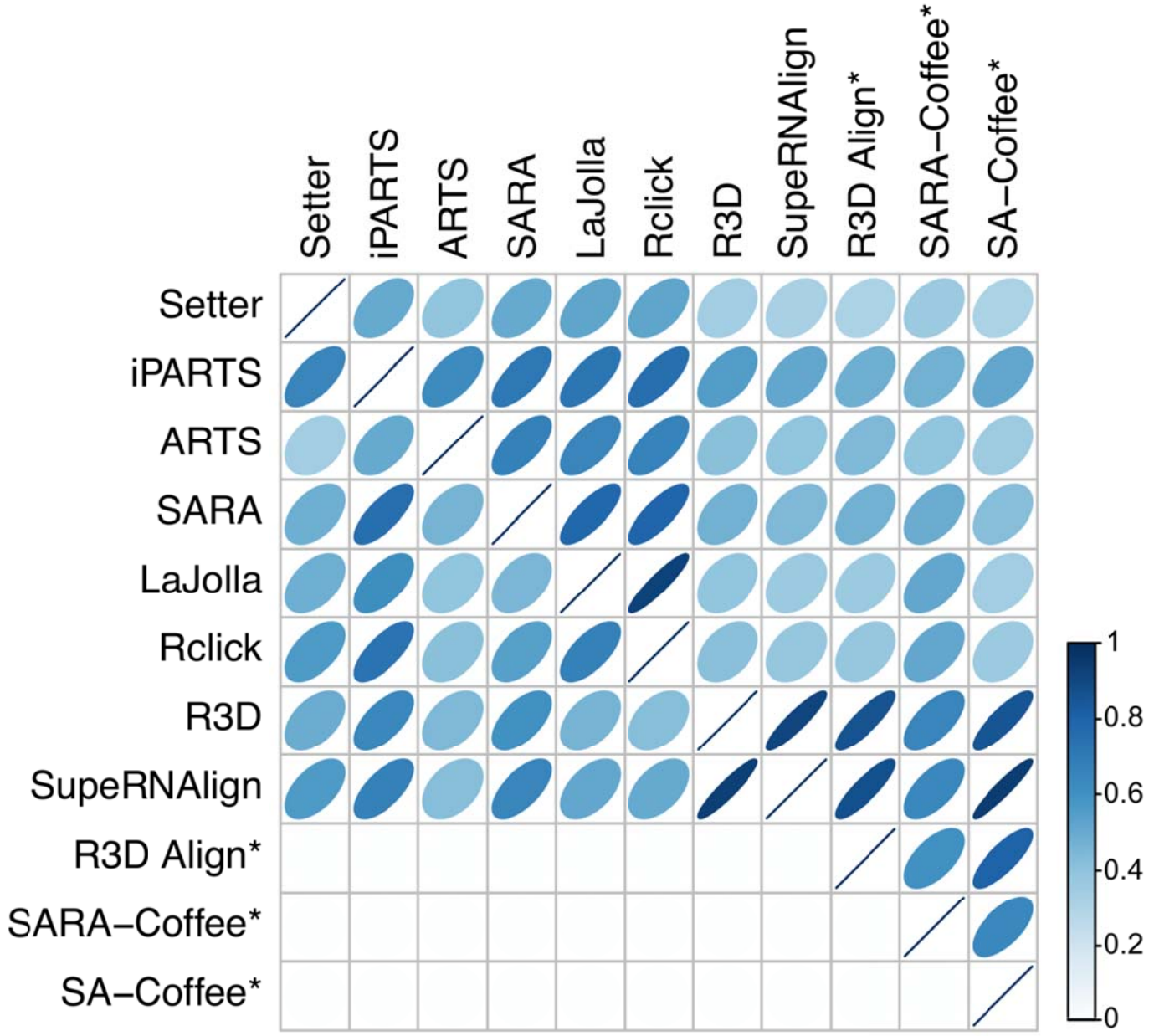
Supplementary Figure 1. A diagram illustrating a graph generated by ClaRNet for the tRNA(Asp) structure (PDB code: 4WJ4_C). Nodes indicate ribonucleotide residues and edges indicate interactions. Residues are labeled with the sequence number. Two clusters of well-connected residues that are less strongly connected to each other are indicated by blue and orange colors, in analogy to Figure 2A.



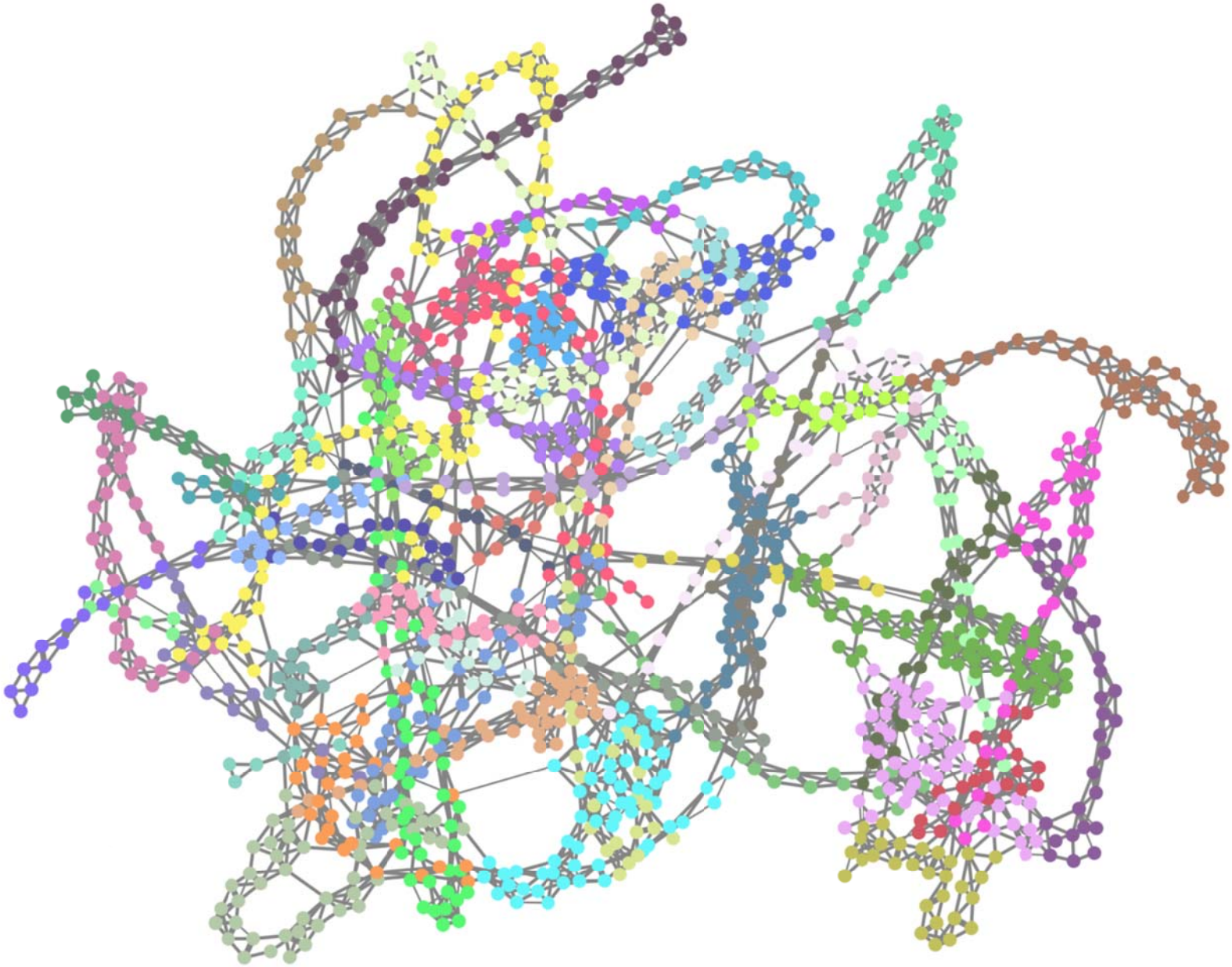
Supplementary Figure 2. Comparison of the sequence alignments for an example pair (1EIY_C and 1QF6_B) between highest-scoring programs and Rfam. Higher similarity of the alignment created by SupeRNAlign-Coffee to the reference alignment can be clearly seen.



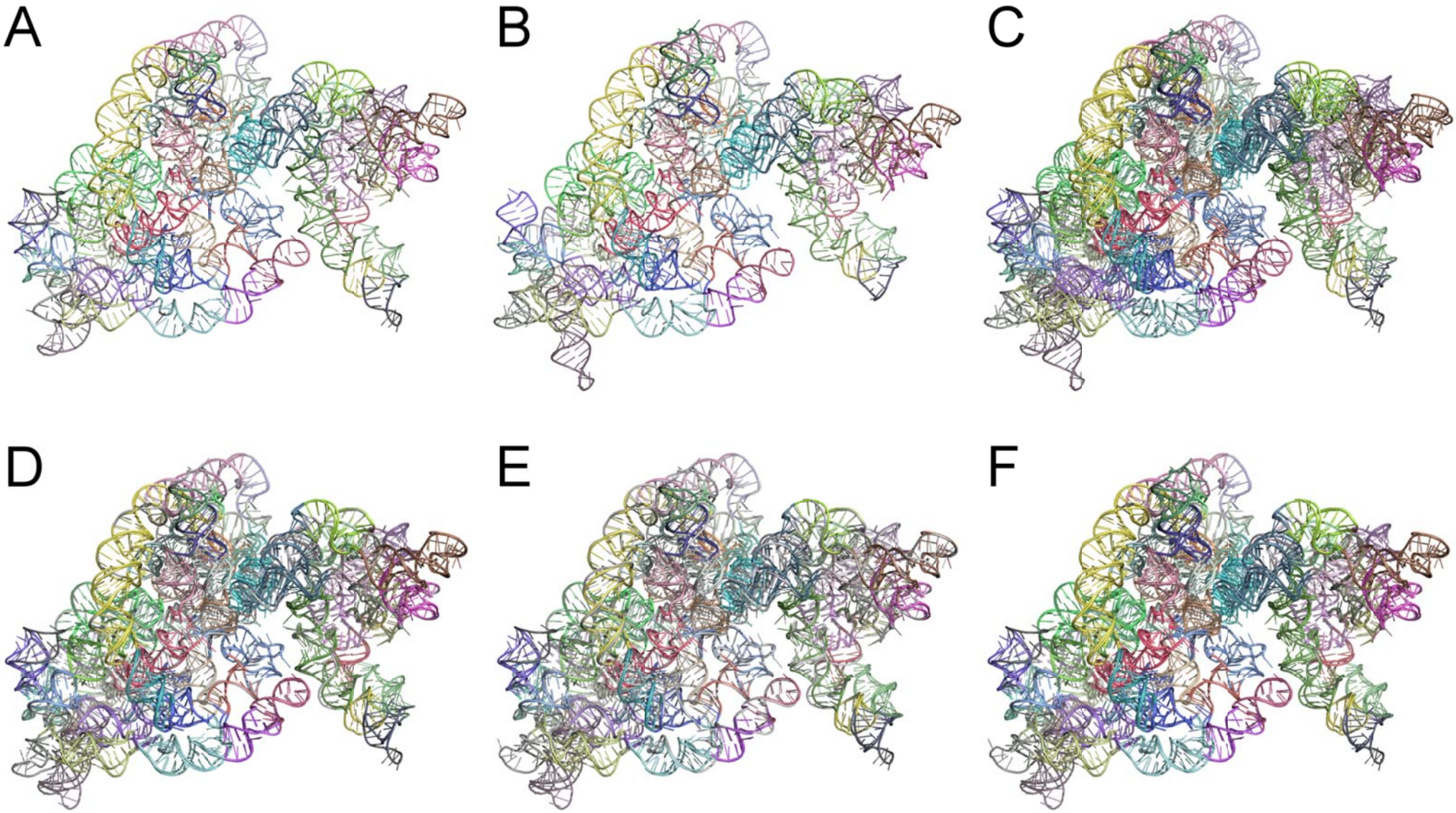
Supplementary Figure 3. Distribution of 3SP scores obtained by the RNA superposition methods. Boxes mark quartiles (Q1, median, Q3); whiskers stretch from 1st to 99th percentile; outliers are shown as dots.



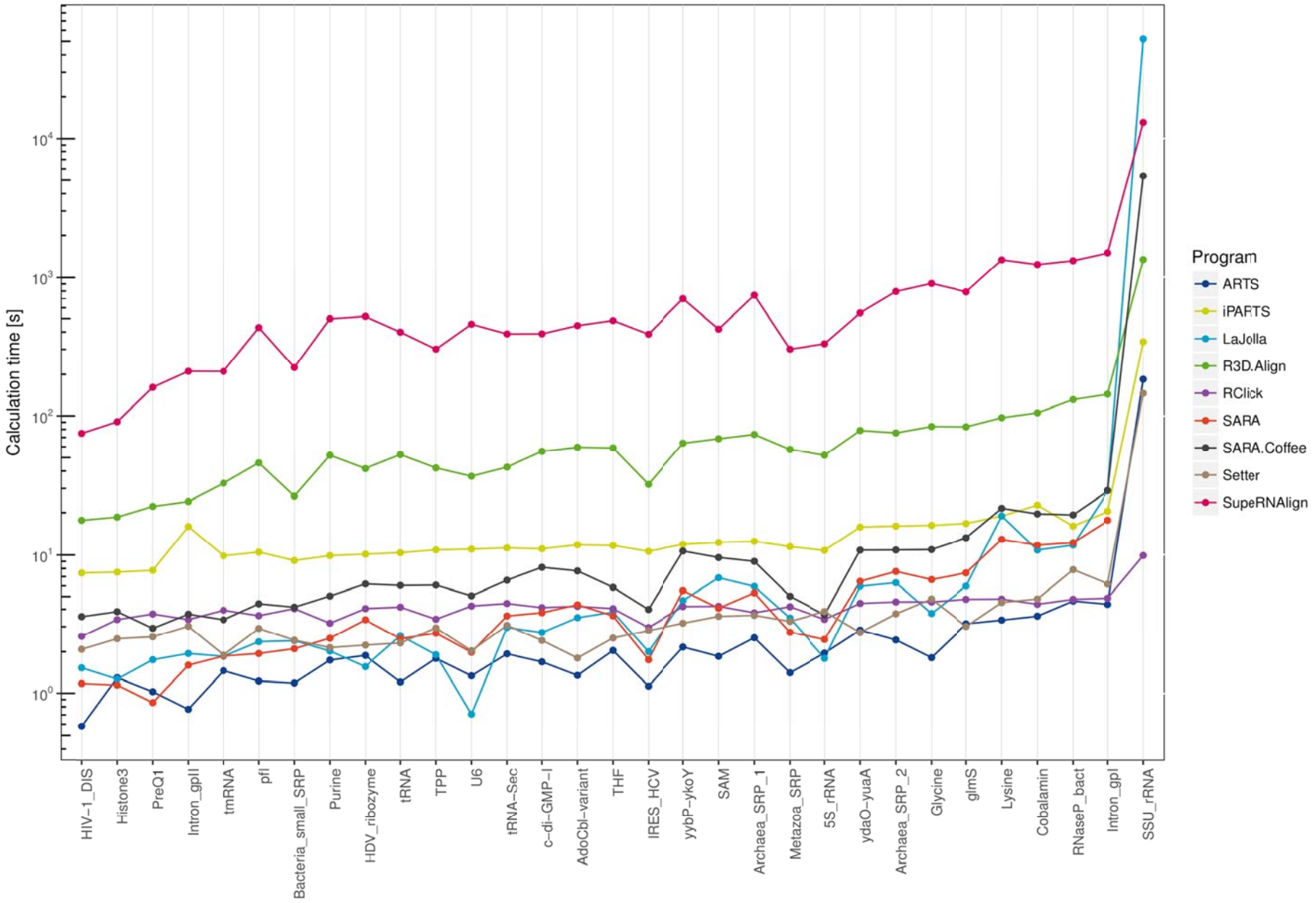
Supplementary Figure 4. Correlations between scores (top-right part of the matrix: sum-of-pairs values, bottom-left part of the matrix, RMSD values) achieved by the benchmarked programs. The darker the ellipse and the higher its ellipticity, the higher Spearman's ρ coefficient.



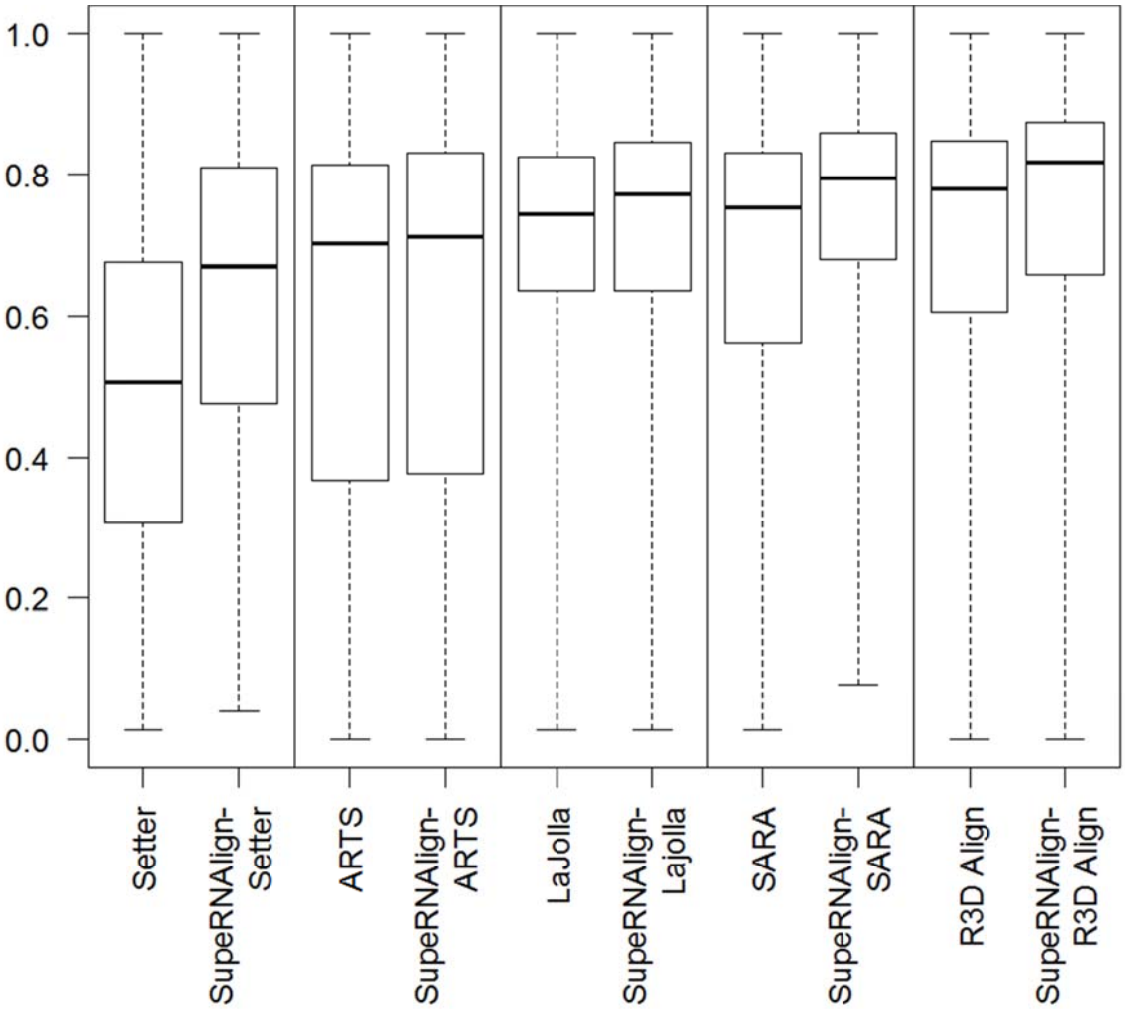
Supplementary Figure 5. A graph generated by ClaRNet for the 16S rRNA structure (PDB code: 2AW7_A), which illustrates that even extremely large and complex structures can be analyzed with the software. Superimposing of this structure on 1FJG_A is depicted in Supplementary Figure 4.



Supplementary Figure 6. Graphical illustration of SupeRNAlign workflow for a pair of 16S rRNA molecules (1FJG_A and 2AW7_A). Colors denote different clusters determined by ClaRNet (see Supplementary Figure 3), or respective regions in the reference structure; frozen fragments are light grey. A) Reference structure (1FJG). B) Aligned structure (2AW7) after clustering by ClaRNet. C) Initial (global) superposition of the two structures. D), E) Intermediate stages; frozen fragments marked light grey. F) Final superposition.



Supplementary Figure 7. Running times (in seconds) for selected structure pairs from each RNA family.



Supplementary Figure 8. The comparison of SPS scores for aligners alone and combined with SupeRNAlign.