Supplementary Materials for
RIblast: An ultrafast RNA-RNA interaction prediction system for comprehensive lncRNA interaction analysis

Tsukasa Fukunaga and Michiaki Hamada

Supplementary Figures

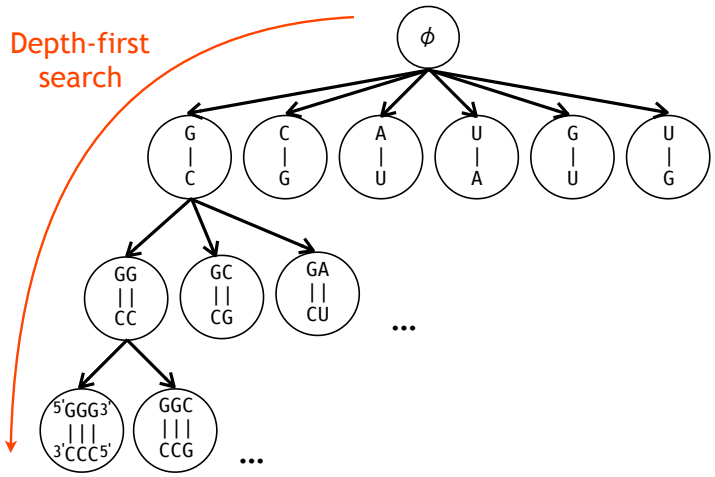

Fig. S1 Schematic illustration of seed search algorithm. RIblast detects seeds using depth-first search.

Algorithm 1 $SeedSearch()$

$S_q \leftarrow$ a query RNA sequence
$S_{db} \leftarrow$ reversed and concatenated target RNA sequences
$SA_q \leftarrow ConstructionSuffixArray(S_q)$
$SA_{db} \leftarrow ConstructionSuffixArray(S_{db})$
$SeedSearchCore(S_q, SA_q, \{\}, S_{db}, SA_{db}, \{\}, 0, |S_q| - 1, 0, |S_{db}| - 1, 0, 0)$

Algorithm 2 $SeedSearchCore(S_q, SA_q, seed_q, S_{db}, SA_{db}, seed_{db}, sp_q, ep_q, sp_{db}, ep_{db}, energy, length)$

if $length < length_{max}$ **then**
  for all c $\in \{\{G,C\},\{C,G\},\{A,U\},\{U,A\},\{G,U\},\{U,G\}\}$ **do**
    $sp, ep \leftarrow SASearchNextCharacter(S_q, SA_q, sp_q, ep_q, c_0, length)$
    $sp', ep' \leftarrow SASearchNextCharacter(S_{db}, SA_{db}, sp_{db}, ep_{db}, c_1, length)$
    if $sp \leq ep$ && $sp' \leq ep'$ **then**
      if $length > 0$ **then**
        $energy \leftarrow energy + CalcStackingEnergy(c_0, c_1, seed_q[\text{length-1}], seed_{db}[\text{length-1}])$
      end if
      if $energy <$ threshold T_1 && $length \geq \delta$ **then**
        store sp, ep, sp', ep'
      else
        $seed_q.PushBack(c_0)$, $seed_{db}.PushBack(c_1)$
        $SeedSearchCore(S_q, SA_q, seed_q, S_{db}, SA_{db}, seed_{db}, sp, ep, sp', ep', energy, length + 1)$
      end if
    end if
  end for
end if
$seed_q.PopBack$, $seed_{db}.PopBack$

Fig. S2 Pseudocode of seed search algorithm. The $ConstructionSuffixArray$ function generates a suffix array from an RNA sequence in linear-time order to sequence length. $length_{max}$ is the max seed length. The $SASearchNextCharacter$ function returns the indices of the new extended string on a suffix array by binary search. The $CalcStackingEnergy$ function returns the energies of stack consisting of two base pairs. $seed_q$ and $seed_{db}$ are temporary seeds on query and database, respectively. If extended strings are detected, the hybridization energies is smaller than T_1 and the length is δ and more, the indices of the strings on suffix arrays are stored.

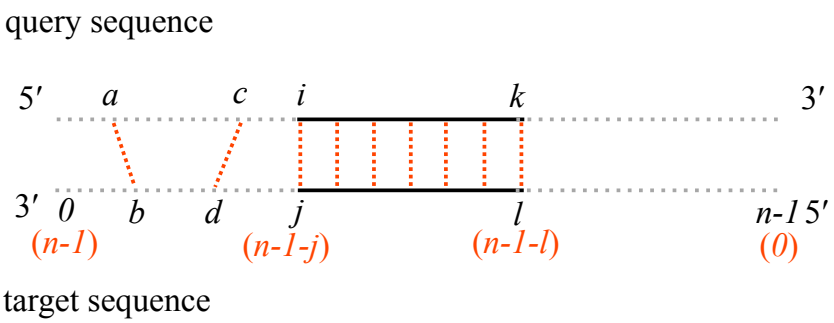

Fig. S3 Schematic illustration of the calculation of interaction energy in gapped extension. The pair of black solid lines represent an interaction after gapless extension. The gray dotted lines are RNA sequences that does not form inter-molecular base pair in gapless extension. The red dotted lines represent inter-molecular base pairs. The black indices are the indices for calculation of hybridization energy, and the red indices are the indices for obtaining accessible energy. Note that two indices are different in target sequences because target sequences are reversed for the seed search after calculation of accessible energy.

Supplementary Tables

Table S1 Dependence of the MCC values of bacterial sRNA basepair prediction on T_1 and X parameters when the energy parameter was Turner parameter

	X						
T_1	10	12	14	16	18	20	22
-13.0	0.53	0.54	0.54	0.55	0.55	0.55	0.55
-12.5	0.54	0.56	0.56	0.58	0.58	0.58	0.58
-12.0	0.55	0.56	0.57	0.59	0.59	0.58	0.59
-11.5	0.53	0.56	0.56	0.58	0.58	0.58	0.57
-11.0	0.54	0.58	0.58	0.60	0.60	0.60	0.59
-10.5	0.55	0.59	0.58	0.59	0.59	0.59	0.58
-10.0	0.56	0.58	0.58	0.59	**0.61**	0.60	0.60
-9.5	0.56	0.58	0.58	0.59	0.60	0.59	0.59
-9.0	0.57	0.58	0.57	0.58	0.60	0.59	0.60

The row and the column indicates X values and T_1 values, respectively. T_1 is a threshold energy for score-based seed detection, and X is a threshold length for extension termination. The bold values mean the best MCC score.

Table S2 Dependence of the MCC values of bacterial sRNA basepair prediction on T_1 and X parameters when the energy parameter was Andronescu parameter

	X						
T_1	X=10	12	14	16	18	20	22
-6.5	0.64	0.65	0.65	0.66	0.67	0.66	0.67
-6.25	0.63	0.64	0.64	0.65	0.66	0.66	0.66
-6.0	0.65	0.66	0.65	**0.67**	0.67	0.67	0.65
-5.75	0.65	0.66	0.65	0.67	0.67	0.67	0.65
-5.5	0.65	0.66	0.65	0.66	0.67	0.67	0.65
-5.25	0.65	0.66	0.65	0.66	0.67	0.67	0.65
-5.0	0.65	0.66	0.65	0.66	0.67	0.67	0.65
-4.75	0.65	0.66	0.65	0.66	0.67	0.67	0.65
-4.5	0.65	0.66	0.65	0.66	0.67	0.67	0.65

The row and the column indicates X values and T_1 values, respectively. T_1 is a threshold energy for score-based seed detection, and X is a threshold length for extension termination. The bold values mean the best MCC score.

Table S3 Dependence of the MCC values of bacterial sRNA basepair prediction on Y parameters

Y	Turner parameter	Andronescu parameter
3	0.61	0.67
4	0.61	0.67
5	0.61	0.67
6	0.61	0.67
7	0.60	0.67

The row and the column indicates the kinds of energy parameter and Y values, respectively. Y is a threshold length for extension termination in gapless extension step.

Table S4 Dependence of the MCC values of bacterial sRNA basepair prediction on W parameters

W	Turner parameter	Andronescu parameter
30	0.60	0.59
50	0.57	0.64
70	0.61	0.67
100	0.61	0.63
150	0.59	0.63

The row and the column indicates the kinds of energy parameter and W values, respectively. W is the constraint of maximal distance between the bases that may form base pairs.

Table S5 Dependence of the MCC values of bacterial sRNA basepair prediction on W parameters

δ	Turner parameter	Andronescu parameter
3	0.58	0.65
4	0.60	0.66
5	0.61	0.67
6	0.57	0.64
7	0.59	0.61

The row and the column indicates the kinds of energy parameter and δ values, respectively. δ is a parameter for accessibility approximation and determination of minimum seed length.

Table S6 TINCR target prediction performance for MINENERGY sorting

	Threshold about the number of interacted segments				
Software	1	2	3	4	5
LAST	**0.583**	**0.587**	**0.578**	0.593	0.570
Terai's pipeline	0.565	0.557	0.557	0.579	**0.577**
RIblast (Turner)	0.573	0.560	0.566	0.598	0.559
RIblast (Andronescu)	0.581	0.567	0.572	**0.600**	0.574

The row is a threshold about the number of interacted segments, which defines the positive dataset. The column shows each software. The bold values mean the best AUROC score.

Table S7 TINCR target prediction performance of RIblast with Turner parameter for SUMENERGY sorting

	Threshold about the number of interacted segments				
interaction energy threshold	1	2	3	4	5
0	0.618	0.622	0.630	0.662	0.647
-1	0.618	0.622	0.630	0.662	0.647
-2	0.618	0.622	0.630	0.662	0.647
-3	0.618	0.623	0.631	0.662	0.648
-4	0.618	0.623	0.631	0.663	0.648
-5	0.619	0.625	0.633	0.665	0.650
-6	0.621	0.627	0.635	0.667	0.652
-7	0.623	0.630	0.638	0.671	0.655
-8	0.627	0.634	0.642	0.676	0.659
-9	0.632	0.639	0.646	0.682	0.664
-10	0.639	0.646	0.652	0.688	0.668
-11	0.646	0.653	0.657	0.694	0.674
-12	0.652	0.658	0.661	0.699	0.677
-13	0.657	0.663	0.665	0.703	0.680
-14	0.659	0.666	**0.668**	0.706	0.682
-15	0.661	**0.667**	0.667	**0.707**	**0.682**
-16	**0.662**	**0.667**	0.667	0.706	0.681
-17	0.661	0.665	0.666	0.706	0.679
-18	0.658	0.662	0.663	0.704	0.674
-19	0.656	0.658	0.659	0.700	0.672

The row is a threshold about the number of interacted segments, which defines the positive dataset. The column shows the interaction energy threshold for SUMENEGY sorting. The bold values mean the best AUROC score.

Table S8 TINCR target prediction performance of RIblast with Andronescu parameter for SUMENERGY sorting

	Threshold about the number of interacted segments				
interaction energy threshold	1	2	3	4	5
0	0.613	0.618	0.627	0.659	0.646
-0.5	0.613	0.618	0.627	0.659	0.646
-1.0	0.613	0.618	0.627	0.659	0.646
-1.5	0.613	0.618	0.627	0.659	0.646
-2.0	0.613	0.618	0.627	0.659	0.646
-2.5	0.614	0.619	0.628	0.661	0.647
-3.0	0.615	0.620	0.630	0.663	0.649
-3.5	0.617	0.623	0.632	0.666	0.651
-4.0	0.621	0.627	0.636	0.671	0.654
-4.5	0.627	0.632	0.640	0.676	0.659
-5.0	0.633	0.639	0.645	0.683	0.664
-5.5	0.641	0.646	0.651	0.689	0.668
-6.0	0.648	0.652	0.656	0.694	0.672
-6.5	0.654	0.657	0.660	0.699	0.676
-7.0	0.658	0.662	0.664	0.703	0.680
-7.5	0.662	0.666	0.666	0.705	0.680
-8.0	**0.664**	**0.667**	**0.667**	**0.707**	**0.682**
-8.5	**0.664**	0.666	**0.667**	0.706	0.681
-9.0	**0.664**	0.665	0.665	0.704	0.678
-9.5	0.662	0.662	0.663	0.702	0.676

The row is a threshold about the number of interacted segments, which defines the positive dataset. The column shows the interaction energy threshold for SUMENEGY sorting. The bold values mean the best AUROC score.

Table S9 TINCR target prediction performance of LAST for SUMENERGY sorting

	Threshold about the number of interacted segments				
interaction energy threshold	1	2	3	4	5
-16	0.637	0.631	0.632	0.659	0.636
-17	0.639	0.634	**0.635**	0.660	0.636
-18	0.639	0.635	0.633	**0.662**	**0.638**
-19	0.641	0.637	0.630	0.661	0.636
-20	0.642	**0.641**	**0.635**	0.664	0.637
-21	**0.643**	**0.641**	0.629	0.660	0.635
-22	0.642	0.640	0.624	0.648	0.623
-23	0.642	0.640	0.627	0.648	0.622
-24	0.637	0.636	0.627	0.644	0.625
-25	0.630	0.632	0.618	0.634	0.614
-26	0.625	0.626	0.613	0.638	0.607
-27	0.620	0.623	0.608	0.630	0.599
-28	0.612	0.618	0.607	0.624	0.605
-29	0.597	0.610	0.601	0.625	0.611
-30	0.583	0.598	0.584	0.604	0.600

The row is a threshold about the number of interacted segments, which defines the positive dataset. The column shows the interaction energy threshold for SUMENEGY sorting. The bold values mean the best AUROC score.

Table S10 Dependence of the AUROC scores and the calculation time of TINCR target prediction on T_2 parameter when the energy parameter was Turner parameter

		Threshold about the number of interacted segments				
T_2	Calculation time (s)	1	2	3	4	5
0	14085	0.662	0.667	0.667	0.706	0.681
-2	12361	0.661	0.667	0.667	0.706	0.681
-4	7817	0.660	0.665	0.665	0.705	0.680
-6	5862	0.658	0.663	0.662	0.703	0.680
-8	3778	0.655	0.660	0.660	0.701	0.677
-10	3896	0.654	0.658	0.661	0.701	0.677
-12	1884	0.652	0.654	0.659	0.697	0.673
-14	1204	0.648	0.648	0.654	0.693	0.668
-16	1073	0.638	0.636	0.637	0.672	0.656

The row is the calculation time and a threshold about the number of interacted segments, which defines the positive dataset. The column shows T_2 values, which was the exclusion threshold energy after gapless extension for speeding up the computation.

Table S11 Dependence of the AUROC scores and the calculation time of TINCR target prediction on T_2 parameter when the energy parameter was Andronescu parameter

		Threshold about the number of interacted segments				
T_2	Calculation time (s)	1	2	3	4	5
0	11669	0.664	0.667	0.667	0.707	0.682
-1	10597	0.663	0.667	0.667	0.707	0.682
-2	8483	0.663	0.667	0.667	0.707	0.682
-3	5771	0.662	0.666	0.667	0.707	0.683
-4	3486	0.660	0.663	0.665	0.706	0.681
-5	2145	0.656	0.659	0.662	0.701	0.676
-6	1301	0.653	0.656	0.660	0.697	0.674
-7	1591	0.650	0.654	0.657	0.695	0.672
-8	803	0.649	0.651	0.655	0.690	0.668

The row is the calculation time and a threshold about the number of interacted segments, which defines the positive dataset. The column shows T_2 values, which was the exclusion threshold energy after gapless extension for speeding up the computation.