Supporting Information

LinearFold: Linear-Time Approximate RNA Folding by 5'-to-3' Dynamic Programming and Beam Search

Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David Hendrix, and David H. Mathews

A Extra Definitions

In Section 2.1, we sketched the definition of the set of allowed pseudoknot-free secondary structures

$$\mathcal{Y}(\mathbf{x}) = \left\{\mathbf{y} \in \{\texttt{.}, \texttt{(}, \texttt{)}\}^{|\mathbf{x}|} \mid \text{balanced}(\mathbf{y}), \text{valid}(\mathbf{x}, \text{pairs}(\mathbf{y}))\right\}$$

Here we complete it. First we denote $\text{depth}(\mathbf{y}) = \sum_i \left(\mathbb{1}[y_i = \texttt{(}] - \mathbb{1}[y_i = \texttt{)}]\right)$ to be the difference in counts between "(" and ")" in $\mathbf{y}$, and then $\text{balanced}(\mathbf{y})$ is true iff.:

$$\forall i, \text{depth}(y_1 ... y_i) \geq 0; \text{ and } \text{depth}(\mathbf{y}) = 0.$$

We next define the set of pairs in $\mathbf{y}$:

$$\text{pairs}(\mathbf{y}) = \{(i, j) \mid y_i = \texttt{(}, y_j = \texttt{)}, \text{balanced}(y_i ... y_j)\}$$

and $\text{valid}(\mathbf{x}, S)$ checks if all pairs in set S are valid for $\mathbf{x}$, i.e., it returns true iff.:

$$\forall (i, j) \in S, x_i x_j \in \{\text{CG, GC, AU, UA, GU, UG}\}$$

We also define $\text{unpaired}(\mathbf{y}) = \{i \mid y_i = \texttt{.}\}$ to be set of unpaired indices in $\mathbf{y}$.

B Actual Scoring Functions

The actual scoring functions used by CONTRAfold, RNAfold, and our LinearFold decompose into individual loops:

$$\begin{aligned} sc_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = & \sum_{(i,j) \in \text{hairpin\_loops}(\mathbf{y})} sc^{\text{H}}_{\mathbf{w}}(\mathbf{x}, i, j) + \sum_{(i,j,k,l) \in \text{single\_loops}(\mathbf{y})} sc^{\text{S}}_{\mathbf{w}}(\mathbf{x}, i, j, k, l) \\ & + \sum_{m \in \text{multi\_loops}(\mathbf{y})} sc^{\text{M}}_{\mathbf{w}}(\mathbf{x}, m) + \sum_{(i,j) \in \text{external\_loops}(\mathbf{y})} sc^{\text{E}}_{\mathbf{w}}(\mathbf{x}, i, j). \end{aligned} \tag{3}$$

where $sc^{\text{H}}_{\mathbf{w}}(\mathbf{x}, \cdot, \cdot)$, $sc^{\text{S}}_{\mathbf{w}}(\mathbf{x}, \cdot, \cdot, \cdot, \cdot)$, $sc^{\text{M}}_{\mathbf{w}}(\mathbf{x}, \cdot)$, $sc^{\text{E}}_{\mathbf{w}}(\mathbf{x}, \cdot, \cdot)$ are scores of hairpin loop, single loop (including bulge and internal loop and stacking), multiloop and external loop, respectively. Multiloop score can be further decomposed into each adjacent base pair $(i, j) \in m$:

$$sc^{\text{M}}_{\mathbf{w}}(\mathbf{x}, m) = w^{\text{multi}}_{\text{base}} + w^{\text{multi}}_{\text{unpair}} \cdot |\text{unpaired}(m)| + \sum_{(i,j) \in m} w^{\text{multi}}_{\text{bp}}(\mathbf{x}, i, j) \tag{4}$$

For example, if $\mathbf{y}$ = `.((...))(((...)))..`, then $\text{multi\_loops}(\mathbf{y})$ is a singleton-set containing $m = ((2, 16), (4, 8), (9, 15))$ with $\text{unpaired}(m) = \{3\}$, $\text{hairpin\_loops}(\mathbf{y}) = \{(4, 8), (10, 14)\}$, $\text{single\_loops}(\mathbf{y}) = \{(9, 10, 14, 15)\}$, and $\text{external\_loops}(\mathbf{y}) = \{(0, 2), (16, 17)\}$.

The thermodynamic model in Vienna RNAfold scores each type of loop using several feature templates such as hairpin/bulge/internal loop lengths, terminal mismatches, helix stacking, helix closing, etc. The machine-learned model in CONTRAfold replaces energies in the above framework with model weights learned from data. Figure SI 6 implement LinearFold for this scoring function.

C Extra Results Tables and Figures

Tables SI 1 & SI 2 detail the accuracy results (PPV & Sensitivity) from Figure 4. We choose the ArchiveII dataset (Sloma and Mathews, 2016), a diverse set of over 3,000 RNA sequences with known secondary structures. But since the current CONTRAfold machine-learned model (v2.02) is trained on the S-Processed dataset (Andronescu *et al.*, 2007) we removed those sequences that appeared in the S-Processed dataset. The resulting dataset we used contains 2,889 sequences over 9 families, with an average length of 222.2 *nt*.

We sample RNAcentral dataset by evenly splitting the length range from $1,000$ to $244,296$ (the longest sequence) into 30 bins by log-scale, and for each bin randomly select one sequence.

Due to the uncertainty of base-pair matches existing in comparative analysis and the fact that there is fluctuation in base pairing at equilibrium, we consider a base pair to be correctly predicted if it is also displaced by one nucleotide on a strand (Sloma and Mathews, 2016). Generally, if a pair (i, j) is in the predicted structure, we consider it a correct prediction if one of (i, j), $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, $(i, j+1)$ is in the ground truth structure. We also report the accuracy using exact base pair matching instead of this method, in Table SI 2. Both sensitivity and PPV are reported. Generally, if $\hat{\mathbf{y}}$ is the predicted structure and $\mathbf{y}^*$ is the ground truth, we have Sensitivity $= \frac{|\text{pairs}(\hat{\mathbf{y}}) \cap \text{pairs}(\mathbf{y}^*)|}{|\text{pairs}(\mathbf{y}^*)|}$, and PPV $= \frac{|\text{pairs}(\hat{\mathbf{y}}) \cap \text{pairs}(\mathbf{y}^*)|}{|\text{pairs}(\hat{\mathbf{y}})|}$.

The following Figure details the impact of beam size on the number of pairs predicted. Figure SI 2A plots the number of pairs predicted (per nucleotide) with varying beam size, compared with ground truth (both with and without the pseudoknotted pairs). It shows that (a) there are on average 0.2776 pairs per nucleotide in this dataset (meaning about 55.5% of all nucleotides are paired) and 7.6% pairs are pseudoknotted; (b) ViennaRNA tends to overpredict, while CONTRAfold tends to underpredict; (c) our algorithm predicts more pairs with larger beam size; and (d) with the default beam size, it predicts almost the same amounts of pairs as the baselines (only 0.0002 and 0.0012 pairs less per nucleotide, respectively). This is also confirmed by Fig. SI 2B–C.

	# of seqs		avg.	CONTRAfold ♣		LinearFold-C ♣		CONTRAfold		LinearFold-C		Vienna RNAfold		LinearFold-V	
Family	total	used	length	PPV	sens	ΔPPV	Δsens	PPV	sens	ΔPPV	Δsens	PPV	sens	ΔPPV	Δsens
tRNA	557	74	77.3	68.89	70.54	+0.00	+0.00	69.05	70.54	+0.00	+0.00	63.51	72.92	+0.24	+0.19
5S rRNA	1,283	1,125	118.8	73.66	73.74	+0.00	+0.00	75.52	75.61	+0.00	+0.00	59.55	65.96	+0.03	+0.04
SRP	928	886	186.1	62.73	62.41	-0.07	-0.07	63.27	62.84	-0.04	-0.04	59.91	65.42	†+0.35	+0.27
RNaseP	454	182	344.1	48.91	47.90	-0.22	†-0.54	48.96	47.67	-0.11	-0.14	47.28	55.15	+0.12	-0.07
tmRNA	462	462	366	44.88	38.61	†-0.74	‡-0.93	45.74	39.05	†-0.67	‡-0.82	41.47	46.86	‡-0.95	‡-1.02
Group I Intron	98	96	424.9	52.62	50.93	+0.84	†+0.80	52.36	50.64	+0.87	+0.80	46.81	57.68	‡+0.86	†+1.02
telomerase RNA	37	37	444.6	45.39	59.19	-0.05	-0.11	45.62	59.30	-0.05	-0.11	41.47	58.20	+0.05	-0.05
16S rRNA	22	22	1,547.90	41.08	41.77	†+3.56	†+3.09	40.20	41.21	†+3.76	†+3.26	37.23	44.13	†+1.51	+1.59
23S rRNA	5	5	2,927.40	52.47	53.18	†+8.65	†+5.66	48.05	49.61	†+14.03	†+9.86	54.79	62.32	+0.33	+0.16
Overall	3,846	2,889	222.2	54.51	55.36	+1.33	+0.88	54.31	55.16	+1.98	+1.42	50.22	58.74	+0.28	+0.24

Table SI 1. Detailed prediction accuracies in percent, allowing one nucleotide in a pair to be displaced by one position, on the ArchiveII dataset using CONTRAfold MFE, LinearFold-C, Vienna RNAfold and LinearFold-V. This slipping method (Sloma and Mathews, 2016) considers a base pair to be correct if it is slipped by one nucleotide on a strand. ♣ denotes using sharpturn enabled mode (default in CONTRAfold). Statistical significance are marked by †($0.01 \leq p < 0.05$) and ‡($p < 0.01$). Overall, LinearFold-C outperforms CONTRAfold MFE by +1.33/+0.88 in PPV/sensitivity with sharpturn and by +1.98/+ 1.42 in PPV/sensitivity without sharpturn, and LinearFold-V outperforms Vienna RNAfold by +0.28/+0.24 in PPV/sensitivity. Among the nine families, LinearFold-C is significantly better on three (Group I Intron, 16S and 23S rRNAs), and LinearFold-V is significantly better on three (SRP, Group I Intron, and 16S rRNAs). We also report the accuracies using exact base pair match in the next Table.

	# of seqs		avg.	CONTRAfold ♣		LinearFold-C ♣		CONTRAfold		LinearFold-C		Vienna RNAfold		LinearFold-V	
Family	total	used	length	PPV	sens	ΔPPV	Δsens	PPV	sens	ΔPPV	Δsens	PPV	sens	ΔPPV	Δsens
tRNA	557	74	77.3	67.61	69.12	+0.00	+0.00	67.73	69.12	+0.00	+0.00	61.75	70.98	+0.04	-0.07
5S rRNA	1,283	1,125	118.8	70.68	70.70	+0.00	+0.00	72.60	72.59	+0.00	+0.00	57.28	63.35	-0.14	-0.11
SRP	928	886	186.1	59.14	58.61	-0.05	-0.07	59.67	59.02	-0.04	-0.03	56.58	61.55	-0.09	-0.20
RNaseP	454	182	344.1	47.45	46.39	-0.25	†-0.55	47.49	46.15	-0.13	-0.15	45.76	53.28	+0.15	+0.04
tmRNA	462	462	366	42.96	36.94	†-0.81	‡-0.99	43.83	37.38	†-0.72	‡-0.85	39.75	44.90	‡-1.09	‡-1.17
Group I Intron	98	96	424.9	51.21	49.56	+0.80	†+0.75	51.03	49.35	+0.82	+0.74	45.49	56.06	‡+0.81	†+0.97
telomerase RNA	37	37	444.6	43.40	56.58	+0.03	+0.00	43.66	56.72	+0.04	+0.00	39.53	55.40	-0.05	-0.19
16S rRNA	22	22	1,547.90	39.84	40.49	†+3.47	†+2.99	39.01	39.97	†+3.62	†+3.13	35.65	42.26	†+1.33	+1.39
23S rRNA	5	5	2,927.40	50.56	51.24	†+8.51	†+5.60	46.46	47.97	†+13.54	†+9.47	53.20	60.50	+0.07	-0.12
Overall	3,846	2,889	222.2	52.54	53.29	+1.30	+0.86	52.39	53.14	+1.90	+1.37	48.33	56.48	+0.11	+0.06

Table SI 2. The prediction accuracies using exact base-pair matching. Statistical significance are marked by †($0.01 \leq p < 0.05$) and ‡($p < 0.01$). Overall, LinearFold-C outperforms CONTRAfold MFE by +1.30/+0.86 in PPV/sensitivity with sharpturn and by +1.90/+ 1.37 in PPV/sensitivity without sharpturn, and LinearFold-V outperforms Vienna RNAfold by +0.11 PPV and +0.06 sensitivity. Among the nine families, LinearFold-C is significantly better on three (Group I Intron, 16S and 23S rRNAs), and LinearFold-V is significantly better on two (Group I Intron and 16S rRNAs).

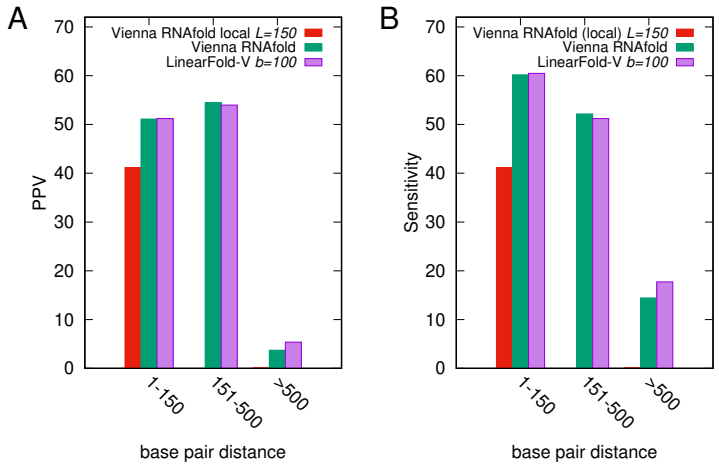

Fig. SI 1. Comparison of LinearFold-V with Vienna RNAfold and its local folding mode in terms of PPV/Sensitivity of base pairs in certain distance ranges across all sequences. LinearFold-V is more accurate in long-range base pairs (500+*nt*) in both PPV and Sensitivity. See Fig. 4C for the corresponding results for LinearFold-C.

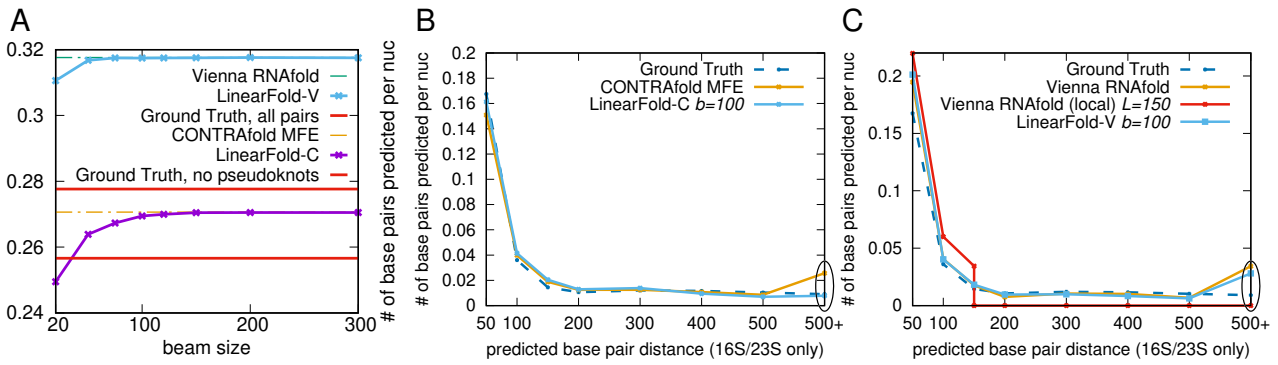

Fig. SI 2. **A**: The number of pairs predicted per nucleotide with varying beam size, comparing these methods and the ground truth (with and without pseudoknots (PK)); **B** and **C**: Length distributions of the predicted base pairs using different methods, on the 16S/23S rRNAs in the ArchiveII dataset. Here we plot the number of both predicted and ground truth base pairs (including pseudoknots) in each of the following ranges: $(0, 50]$, $(50, 100]$, ... $(400, 500)$, $[500, \infty)$. This figure shows that LinearFold-C produces almost the same length distributions with the ground truth, while CONTRAfold severely overpredicts base pairs longer than 500*nt* apart. Both ViennaRNA and LinearFold-V overpredict in that range, but LinearFold-V is less severe. In C, we also reconfirm the limitation of local folding which does not output any long-range pairs.

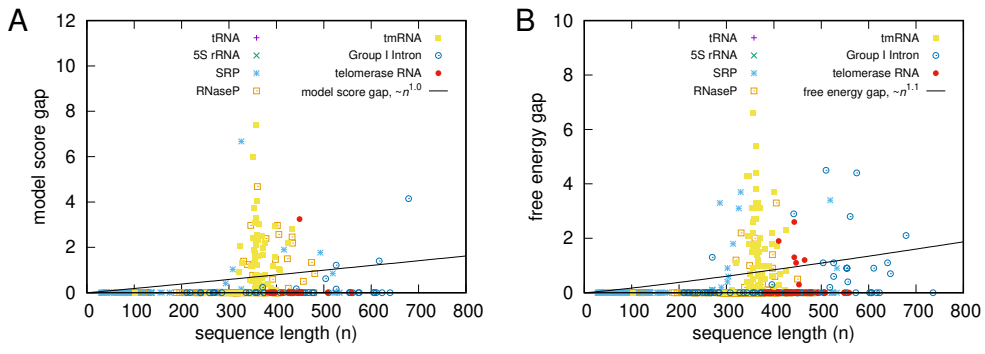

Fig. SI 3. Close-ups for Fig. 5 (search error against sequence length) for for short sequences. A: LinearFold-C vs. CONTRAfold MFE; B: LinearFold-V vs. Vienna RNAfold. Again, tmRNA is the outlier with disproportionally severe search errors, which can explain the slightly worse accuracies of LinearFold on tmRNA in Fig. 4A. Sequences of 250*nt* or less have no search errors (i.e., LinearFold with $b = 100$ is exact for $n \leq 250$).

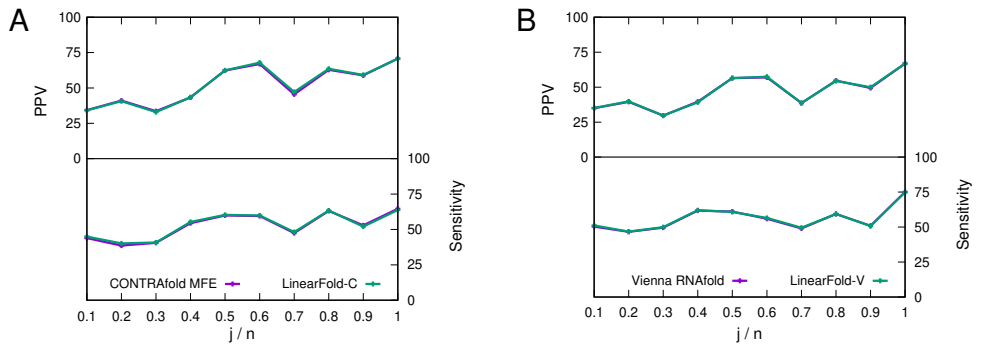

Fig. SI 4. PPV/Sensitivity for all pairs (i, j) as a function of j/n where n is the sequence length, i.e., the "proportional distance" of a pair's right nucleotide to the 5'-end. We bin j/n by $(0, 0.1]$, $(0.1, 0.2]$,..., $(0.9, 1.0]$. In general, LinearFold performs very similarly to the baselines, and even though it scans 5'-to-3', the accuracy does not degrade towards the 3'-end.

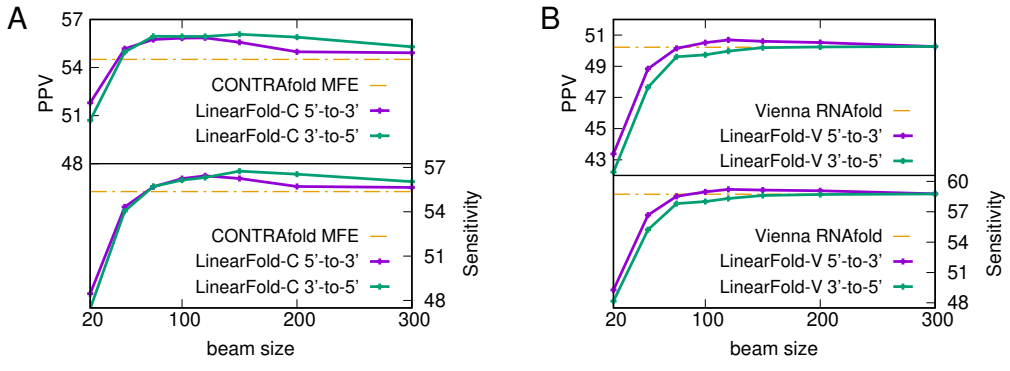

Fig. SI 5. Comparing 5'-to-3' and 3'-to-5' versions of LinearFold. The physical model (B) seems to prefer the default 5'-to-3' order.

D Deductive System for the Actual Systems

The following Figure sketches the deductive system for the actual LinearFold system based on the real scoring functions in Section B. For more implementation details, we refer the readers to our released source code at `https://github.com/LinearFold/LinearFold`.

input $x_1 \dots x_n$

states

$\mathsf{E}\ \langle 0,\ j\rangle : \langle \alpha, s\rangle$ prefix structure

$\mathsf{P}\ \langle i,\ j\rangle : \langle (\alpha), s\rangle$ pair

$\mathsf{H}\ \langle i,\ j\rangle : \langle (\ldots, s\rangle$ hairpin candidate

$\mathsf{M}_1\ \langle i,\ j\rangle : \langle (\alpha)\,\beta, s\rangle$ one or more pairs

$\mathsf{M}_2\ \langle i,\ j\rangle : \langle (\alpha)\,\beta\,(\gamma), s\rangle$ two or more pairs

$\mathsf{M}\ \langle i,\ j\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\ldots, s\rangle$ multiloop candidate

axiom $\mathsf{E}\ \langle 0,\ 1\rangle : \langle |, 0\rangle$

goal $\mathsf{E}\ \langle 0,\ n{+}1\rangle : \langle \alpha, \_\rangle$

push
$$\frac{\mathsf{E}\langle 0,\ j\rangle : \langle \alpha, s\rangle}{\mathsf{H}\langle j,\ \mathrm{next}(j,j)\rangle : \langle (\ldots, 0\rangle}$$
$\mathrm{next}(i,j) \triangleq \min\{k \mid k > j,\ (x_i, x_k)\ \text{match}\}$

Hjump
$$\frac{\mathsf{H}\langle i,\ j\rangle : \langle (\ldots, s\rangle}{\mathsf{H}\langle i,\ \mathrm{next}(i,j)\rangle : \langle (\ldots\ldots, s\rangle}$$

skip
$$\frac{\mathsf{E}\langle 0,\ j\rangle : \langle \alpha, s\rangle}{\mathsf{E}\langle 0,\ j{+}1\rangle : \langle \alpha\,., s{+}sc^{\mathrm{E}}_{\mathbf{w}}(\mathbf{x}, j, j+1)\rangle}$$
$$\frac{\mathsf{M}_1\langle i,\ j\rangle : \langle (\alpha)\,\beta, s\rangle}{\mathsf{M}_1\langle i,\ j{+}1\rangle : \langle (\alpha)\,\beta\,., s{+}w^{\mathrm{multi}}_{\mathrm{unpair}}\rangle}$$

reduce
$$\frac{\mathsf{M}_1\langle k,\ i\rangle : \langle (\alpha)\,\beta, s'\rangle \qquad \mathsf{P}\langle i,\ j\rangle : \langle (\gamma), s\rangle}{\mathsf{M}_2\langle k,\ j\rangle : \langle (\alpha)\,\beta\,(\gamma), s'{+}s{+}w^{\mathrm{multi}}_{\mathrm{bp}}(\mathbf{x}, i, j)\rangle}$$

combine
$$\frac{\mathsf{E}\langle 0,\ i\rangle : \langle \alpha, s'\rangle \qquad \mathsf{P}\langle i,\ j\rangle : \langle (\beta), s\rangle}{\mathsf{E}\langle 0,\ j\rangle : \langle \alpha\,(\beta), s'{+}s{+}\mathrm{sc}^{\mathrm{E}}_{\mathbf{w}}(\mathbf{x}, i, j)\rangle}$$

XtoM$_1$
$$\frac{\mathsf{P}\langle i,\ j\rangle : \langle (\alpha), s\rangle}{\mathsf{M}_1\langle i,\ j\rangle : \langle (\alpha), s{+}w^{\mathrm{multi}}_{\mathrm{bp}}(\mathbf{x}, i, j)\rangle}$$
$$\frac{\mathsf{M}_2\langle i,\ j\rangle : \langle (\alpha)\,\beta\,(\gamma), s\rangle}{\mathsf{M}_1\langle i,\ j\rangle : \langle (\alpha)\,\beta\,(\gamma), s\rangle}$$

Mleft
$$\frac{\mathsf{M}_2\langle i,\ j\rangle : \langle (\alpha)\,\beta\,(\gamma), s\rangle}{\mathsf{M}\langle k,\ \mathrm{next}(k,j)\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\,., s{+}u\cdot w^{\mathrm{multi}}_{\mathrm{unpair}}\rangle}$$
$u = (\mathrm{next}(k,j){-}j){+}(i{-}k{-}1),$
$i{-}k{-}1 \le 30$

Mjump
$$\frac{\mathsf{M}\langle i,\ j\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\ldots, s\rangle}{\mathsf{M}\langle i,\ \mathrm{next}(i,j)\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\ldots\ldots, s{+}u\cdot w^{\mathrm{multi}}_{\mathrm{unpair}}\rangle}$$
$u = \mathrm{next}(i,j){-}j$

hairpin
$$\frac{\mathsf{H}\langle i,\ j\rangle : \langle (\ldots, s\rangle}{\mathsf{P}\langle i,\ j{+}1\rangle : \langle (\ldots), s{+}sc^{\mathrm{H}}_{\mathbf{w}}(\mathbf{x}, i, j)\rangle}$$

singleloop
$$\frac{\mathsf{P}\langle i,\ j\rangle : \langle (\alpha), s\rangle}{\mathsf{P}\langle k,\ l\rangle : \langle (\ldots(\alpha)\ldots), s{+}sc^{\mathrm{S}}_{\mathbf{w}}(\mathbf{x}, i, j, k, l)\rangle}$$
(x_k, x_{l-1}) match, $(l{-}j{-}1){+}(i{-}k{-}1) \le 30$

multiloop
$$\frac{\mathsf{M}\langle i,\ j\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\ldots, s\rangle}{\mathsf{P}\langle i,\ j{+}1\rangle : \langle (\ldots(\alpha)\,\beta\,(\gamma)\ldots), s{+}w^{\mathrm{multi}}_{\mathrm{base}}{+}w^{\mathrm{multi}}_{\mathrm{bp}}(\mathbf{x}, i, j)\rangle}$$

Fig. SI6. The actual deductive system implemented in LinearFold. Shaded substrings are balanced in brackets. Here $sc^{\mathrm{E}}_{\mathbf{w}}(\mathbf{x},\cdot,\cdot)$, $w^{\mathrm{multi}}_{\mathrm{base}}$, $w^{\mathrm{multi}}_{\mathrm{bp}}(\mathbf{x},\cdot,\cdot)$, $w^{\mathrm{multi}}_{\mathrm{unpair}}$, $sc^{\mathrm{S}}_{\mathbf{w}}(\mathbf{x},\cdot,\cdot,\cdot,\cdot)$, $sc^{\mathrm{H}}_{\mathbf{w}}(\mathbf{x},\cdot,\cdot)$ are the various energy or scoring parameters (E stands for external loop, multi for multiloop, S for single loop, and H for hairpin loop). The $\mathrm{next}(i,j)$ returns the next position after x_j that can pair with x_i; this is the "jumping" trick used in CONTRAfold and ViennaRNA. Our final two rules also use this jumping trick in the righthand side loop. The only cubic-time rule is reduce (intermediate step in multiloop), again inspired by CONTRAfold source code.

E Connections between Context-Free Parsing and RNA Folding

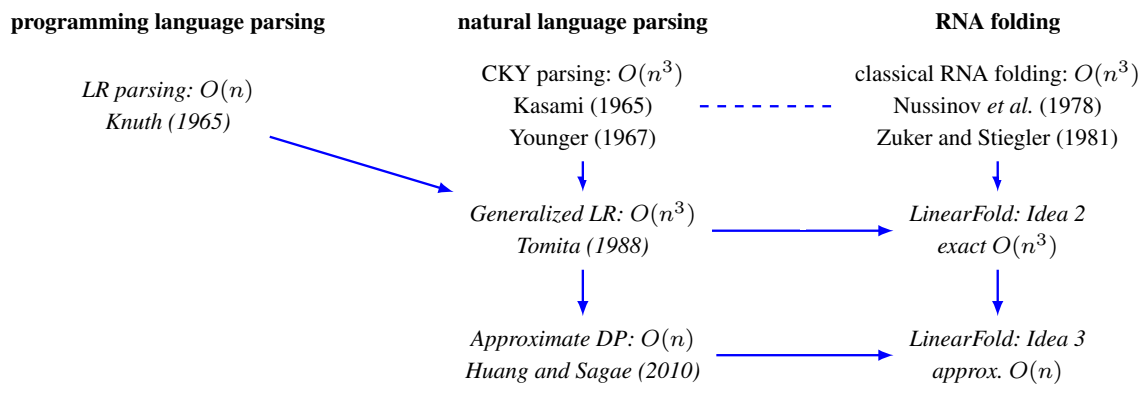

Fig. SI 7. Our work is inspired by incremental parsing algorithms in both programming language theory and computational linguistics. Italic denotes left-to-right algorithms; others are bottom-up. The classical bottom-up $O(n^3)$ algorithms are isomorphic between natural language parsing and RNA folding. Knuth's $O(n)$ LR algorithm works only for a small subset of context-free grammars (CFGs), and Tomita generalizes it to arbitrary CFGs, achieving the alternative, left-to-right, $O(n^3)$ algorithm, which inspires LinearFold Idea 2. Our previous work (Huang and Sagae) modernize and generalize Tomita's algorithm, combining it with beam search to achieve linear runtime, which inspires LinearFold Idea 3.