

Supplementary material - SCARNA:Fast and  
Accurate Structural Alignment of RNA  
Sequences by Matching Fixed-length Stem  
Fragments

## 1 Incremental scores for adjacent matches of stem components

The scores,  $\delta_R(X_i)$ ,  $\delta_f(X_i)$  and  $\delta_e(X_i)$ , are used, when stem components are continuous match. When two stem components match continuously, they must be 1-continuous. So, the match scores,  $\delta_R(X_i)$ ,  $\delta_f(X_i)$  and  $\delta_e(X_i)$ , are computed for incremental region as substitution score, stacking energy and confidence score respectively(see fig.1).



Fig. 1: The match scores for 1-continuous stem component,  $\delta_R(X_i)$ ,  $\delta_f(X_i)$  and  $\delta_e(X_i)$ , are computed for the incremental base pair of stem components colored by blue, because the match scores,  $s(i, j)$ , for the stem components colored by red  $s(i, j)$  have computed as other than the case of 1-continuous match.

## 2 The command line options for other tools

The command line options for alignment tools used by computational experiments on the benchmark dataset of tRNAs(See Section3.1) are listed on Gardner et al. [1]. Others on other benchmark dataset(See Section 3.2 and 4.2) are listed on the following.

Tool	Command
Foldalign2.0	./foldalign -global filename
clustalw	./clustalw filename
carnac	./carnac filename

Table 1:

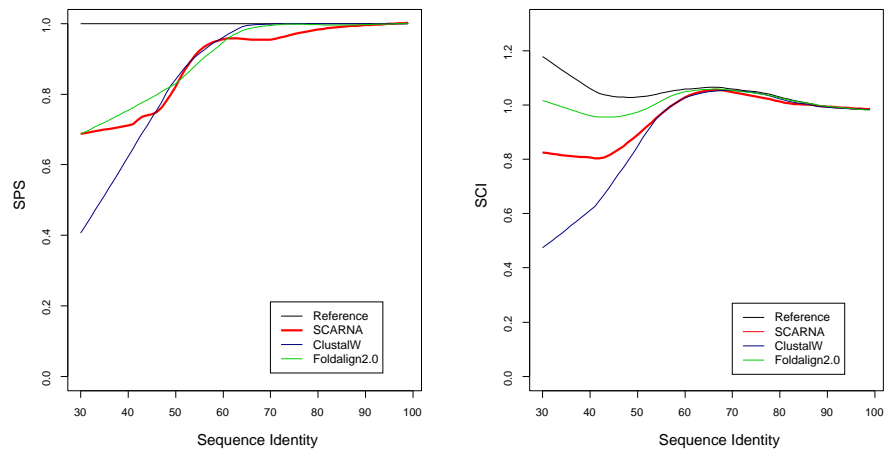


Fig. 2: SPS(left) and SCI(right) as functions of the sequence identity for the dataset of 5S ribosomal RNA which has the Rfam number of RF00001.

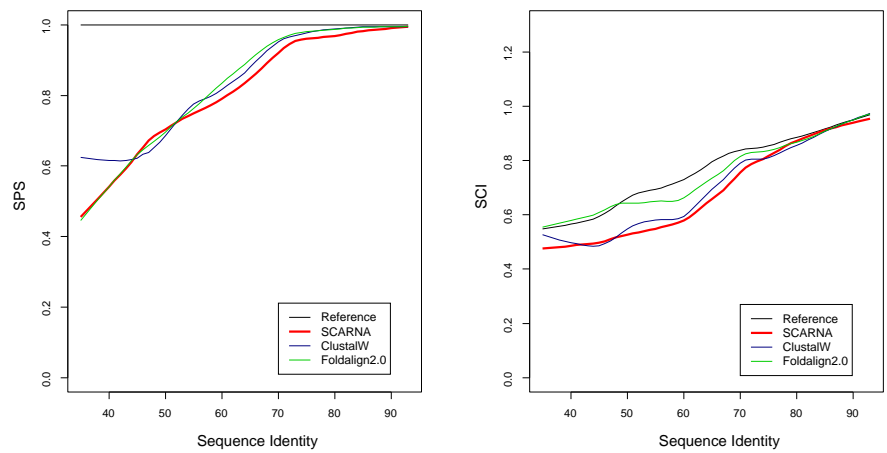


Fig. 3: SPS(left) and SCI(right) as functions of the sequence identity for the dataset of 5.8S ribosomal RNA which has Rfam number of RF00002.

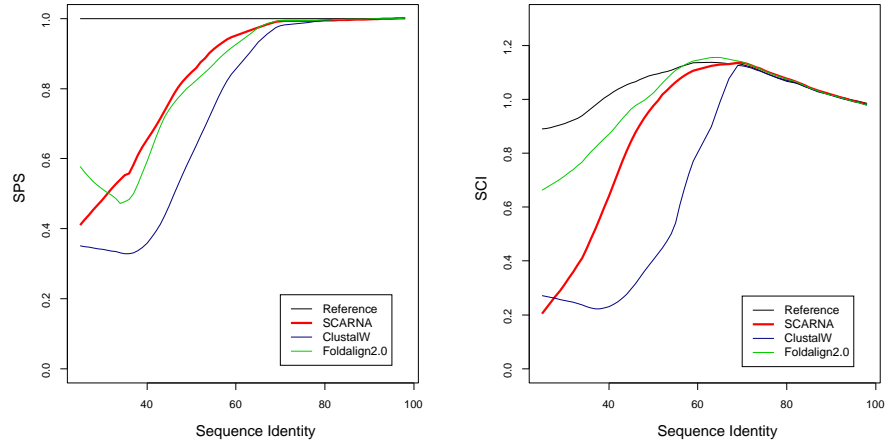


Fig. 4: SPS(left) and SCI(right) as functions of the sequence identity for the dataset of Hammerhead ribozyme which has the Rfam number of RF00008.

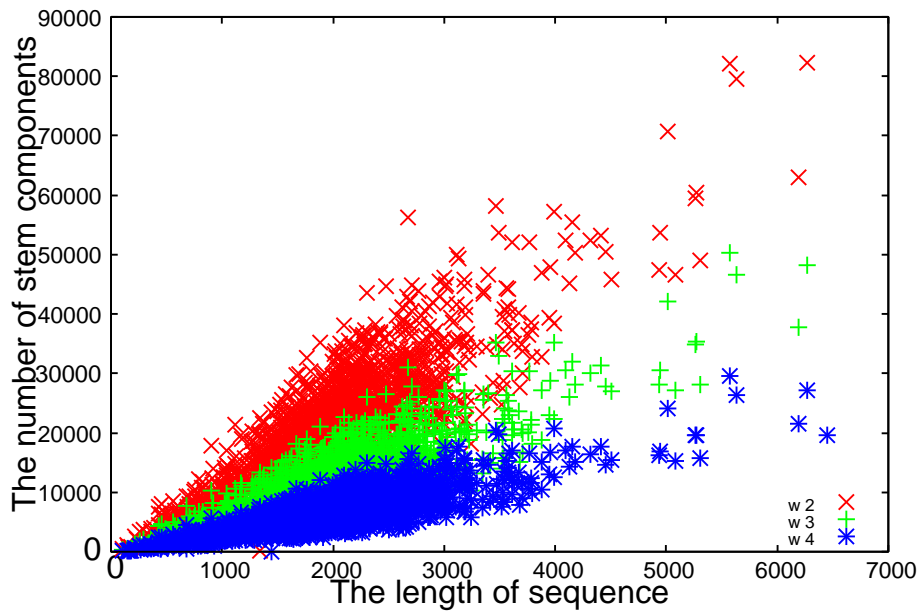


Fig. 5: The number of stem components as function of the length of sequence when base pairs, which have more than 0.0001 probability on base pairing probability matrix, are selected. The data are Non-protein coding transcript cDNA in H-Inv DB. The  $w\#$  indicates the length of stem components.

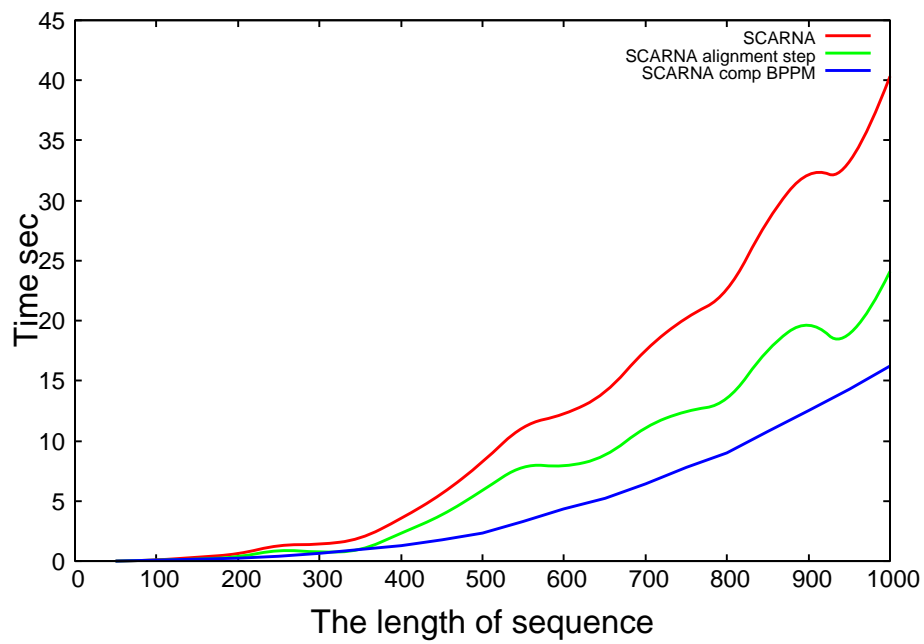


Fig. 6: Execution time of some parts of SCARNA. Red line(SCARNA) shows all execution time of SCARNA. Green line(SCARNA alignment step) shows the execution time of SCS alignment and nucleotide alignment, which is  $O(n^2)$ (n:sequence length) in time. Blue line(SCARNA comp BPPM) shows the computational time of Base Pairing Probability Matrices of two sequences, which is  $O(n^3)$ (n:sequence length) in time.



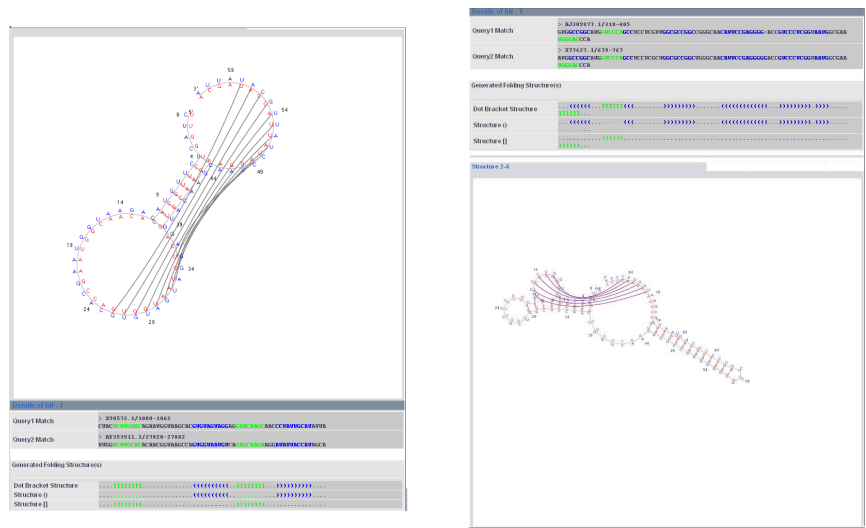


Fig. 9: Web output for Pseudoknot structure prediction of Corona pk3(left) and HDV ribozyme(right).

## References

- [1] P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucl. Acids Res.*, 33(8):2433–2439, 2005.