Supplemental Data

The modified RNAfold program was used with molecule-specific and molecule-independent hairpin loop statistical potentials to predict the secondary structure for sixteen RNA molecular classes. The experimentally-derived energetic parameters for base pair stacks and multi-stem loops were used with both hairpin and internal loop statistical potentials.

Molecule-specific

Using the molecule-specific hairpin loop statistical potentials increased the prediction accuracy relative to the unmodified RNAfold program (TURNER99) for fourteen of the sixteen RNA molecular classes. The increase in accuracy for the fourteen RNA classes ranged from 2% (bacterial SRP) to 25% (HCV IRES); the TPP riboswitch showed no improvement and the SAM riboswitch showed a decrease of 1% **(Fig. S1).**

Molecule-specific internal loop statistical potentials increased prediction accuracy for all sixteen RNA molecular classes. The extent of the increase varied from 8% (TPP riboswitch) to 34% (bacterial 5S and 16S rRNA) **(Fig S1)**. With the exception of tRNA and U1 spliceosomal RNA, all of the RNA molecular classes had larger increases in their prediction accuracy with molecule-specific internal loop statistical potentials compared to molecule-specific hairpin loop statistical potentials **(Supplemental Data, see Excel file Accuracies.xlsx)**.

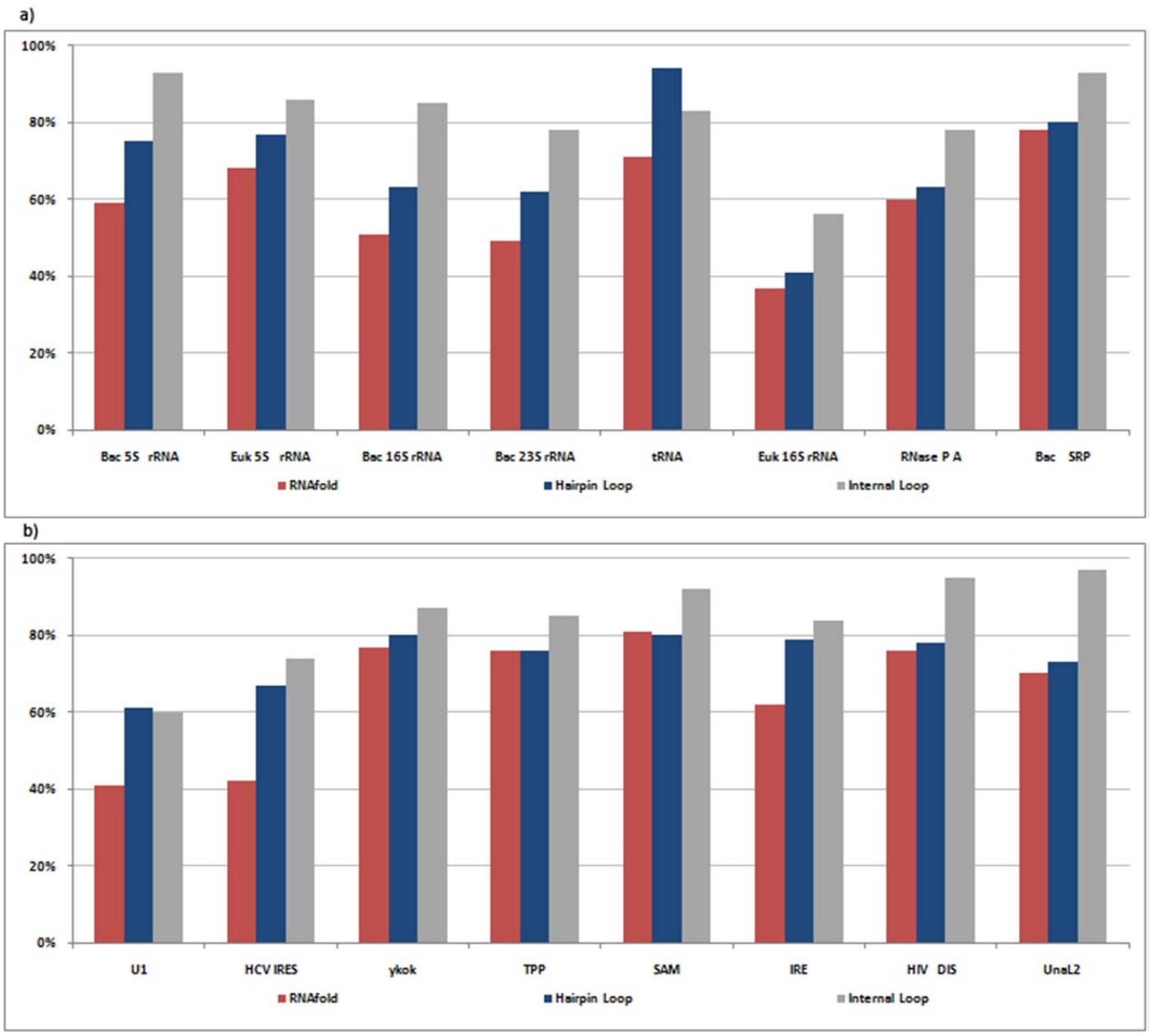

Figure S1: RNA secondary structure prediction accuracies for sixteen RNA molecular classes for RNAfold using 1) unmodified RNAfold, 2) molecule-specific hairpin loop statistical potentials; 3) molecule-specific internal loop statistical potentials. The experimentally derived energetic parameters for base pair stacks and multi-stem loops were used with both hairpin and internal loop statistical potentials. Results for sixteen RNA molecular classes are divided into a) bacterial 5S, 16S, and 23S rRNA, eukaryotic 5S and 16S rRNA, tRNA, RNase P A and bacterial SRP; b) U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader, TPP and SAM riboswitches, iron response element (IRE), HIV type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element.

Molecule-independent

Molecule-independent hairpin loop statistical potentials increased overall prediction accuracy for twelve of the sixteen molecular classes from 1% (eukaryotic nuclear 16S rRNA) to 19% (tRNA) **(Fig. S2)**. Of the four molecular classes with a prediction accuracy for the statistical potential lower than RNAFold (TURNER99): bacterial SRP was 2% lower and ykok leader, TPP and SAM riboswitches were all 3% lower. The prediction accuracies with the statistical potentials for the molecule-independent values were only 0-5% lower than the molecule-specific for twelve of the sixteen molecular classes. Ykok leader and bacterial 23S rRNA were only 6% lower, bacterial 16S rRNA 9% lower and HCV IRES was 13% less.

Fifteen of the sixteen RNA molecular classes had improvements in their prediction accuracy when using the molecule-independent statistical potentials. The increases ranged from 1% (eukaryotic 16S rRNA and RNase P A) to 25% (bacterial 5S rRNA) (**Fig. S2)**. The ykok leader showed no change in accuracy. But the decrease in the prediction accuracy between the molecule-specific vs. independent specific was much greater than for hairpin loops with decreases from 2% to 20%. Similar to the molecule-specific results, all of the RNA molecular classes showed larger increases with molecule-independent internal loop statistical potentials versus the hairpin loop statistical potentials except for eukaryotic 16S rRNA, tRNA and U1 spliceosomal RNA **(Supplemental Data, see Excel file Accuracies.xlsx)**.

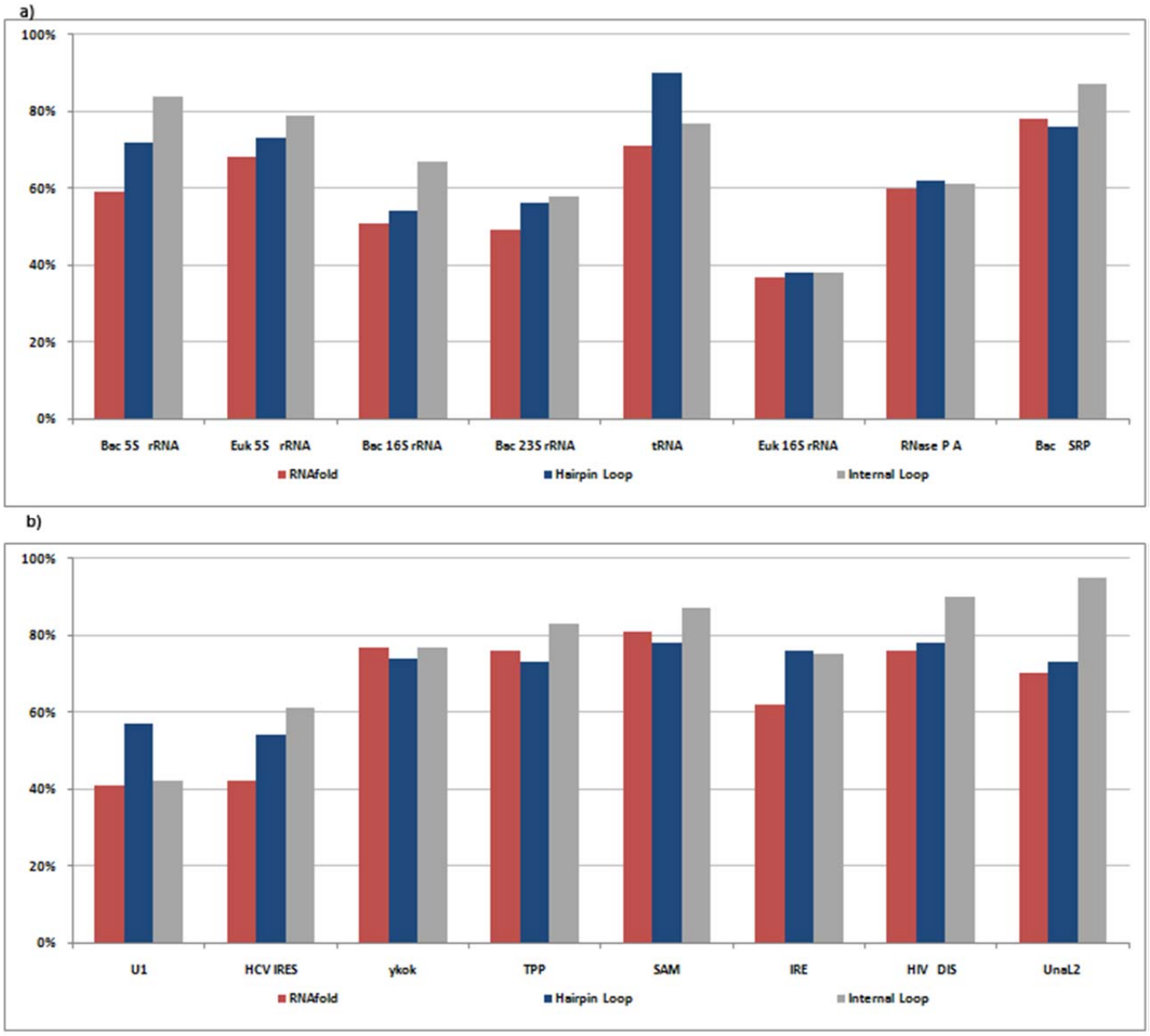

Figure S2: RNA secondary structure prediction accuracies for sixteen RNA molecular classes for RNAfold using 1) unmodified RNAfold, 2) molecule-independent hairpin loop statistical potentials; 3) molecule-independent internal loop statistical potentials. The experimentally derived energetic parameters for base pair stacks and multi-stem loops were used with both the hairpin and internal loop statistical potentials. Results for sixteen RNA molecular classes are divided into a) bacterial 5S rRNA, eukaryotic 5S rRNA, bacterial 16S rRNA, bacterial 23S rRNA, tRNA, eukaryotic 16S rRNA, RNase P A and bacterial SRP; b) U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader, TPP and SAM riboswitches, iron response element (IRE), HIV type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element.

The average standard deviations for the histogram in Figure 2 of the paper for each program for each of the sixteen training RNA molecular classes are shown as error bars in Figures S3. On average, our new statistical potentials had smaller standard deviations than all five of the competing programs **(Supplemental Data, see Excel file Accuracies.xlsx)**.

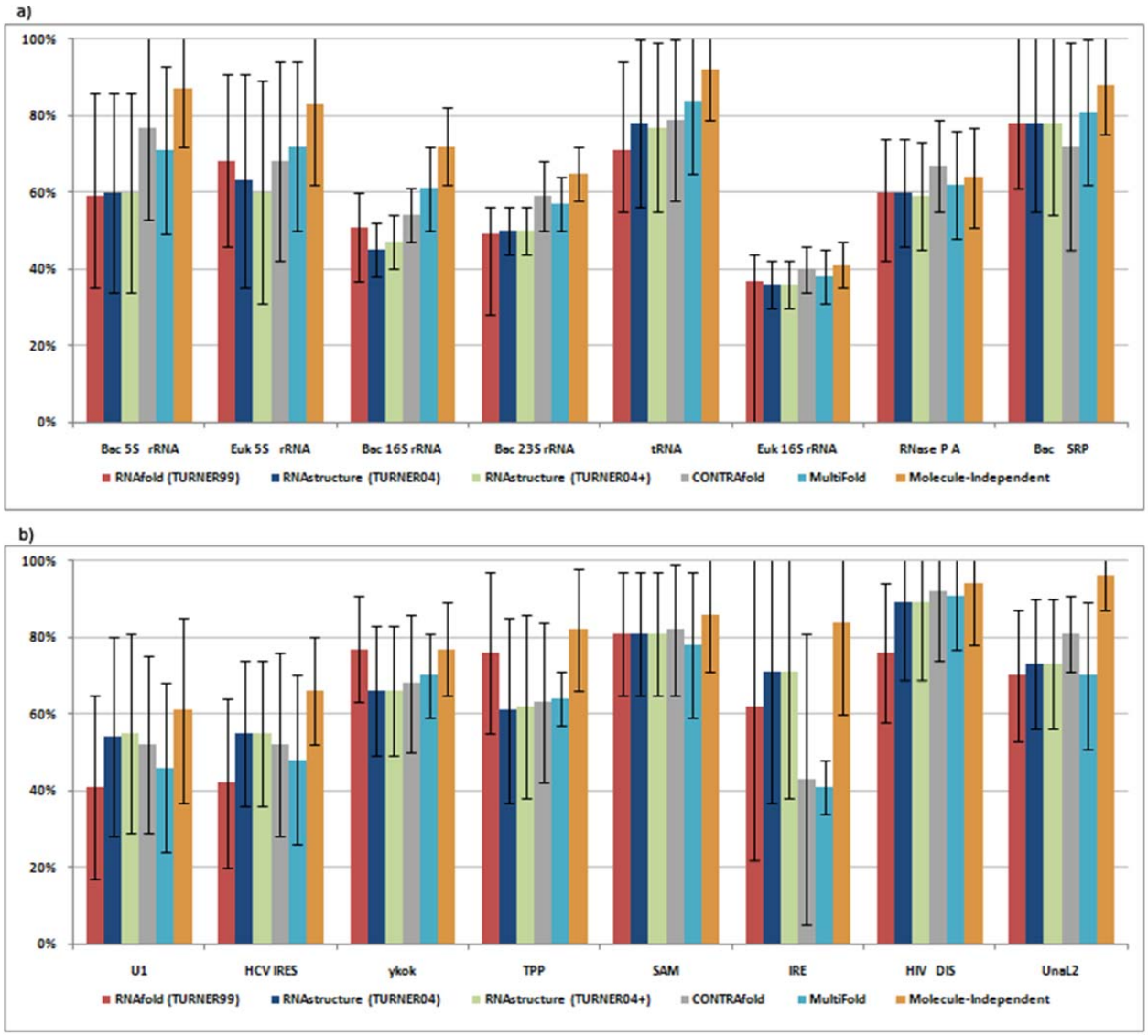

Figure S3: RNA secondary structure prediction accuracies plus standard deviations for five RNA folding programs: RNAfold, RNAstructure (TURNER04 and (TURNER04+), ContraFold, MultiFold and RNAfold using statistical potentials. Results for sixteen RNA molecular classes are divided into a) bacterial 5S rRNA, eukaryotic 5S rRNA, bacterial 16S rRNA, bacterial 23S rRNA, tRNA, eukaryotic 16S rRNA, RNase P A and bacterial SRP; b) U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader, TPP and SAM riboswitches, iron response element (IRE), HIV type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element.

Andronescu et al.[1] followed a standard approach for training and testing; 80% of the sequences were used for training and the remaining 20% for testing. When the statistical potentials are generated and tested following this approach, the results are very similar to those seen when using the full set of sequences for generation and testing.

Overall, the combined molecule-independent statistical potentials still outperformed the other four programs **(Figs. S4a and S4b)**. On average, over the sixteen RNA molecular classes, our statistical potentials scored 15% higher than RNAfold (TURNER99), 14% for RNAstructure (TURNER04), 14% higher for RNAstructure (TURNER04 Plus), 12% for CONTRAfold and 13% for MultiFold. Our statistical potentials outperformed all four programs for all sixteen RNA molecular classes with the exception of the ykok leader RNA where RNAfold (TURNER99) matched our score and RNase P A where CONTRAfold scored 2% higher. The difference in accuracy between our statistical potentials and the program with the best results for a given molecule ranged from -2% (RNase P A) to 16% (IRE) **(Figs. S4a and S4b)**. On average, our statistical potentials outperformed the program with the best results for a given RNA molecule by 6% **(Supplemental Data, see Excel file Accuracies.xlsx)**.

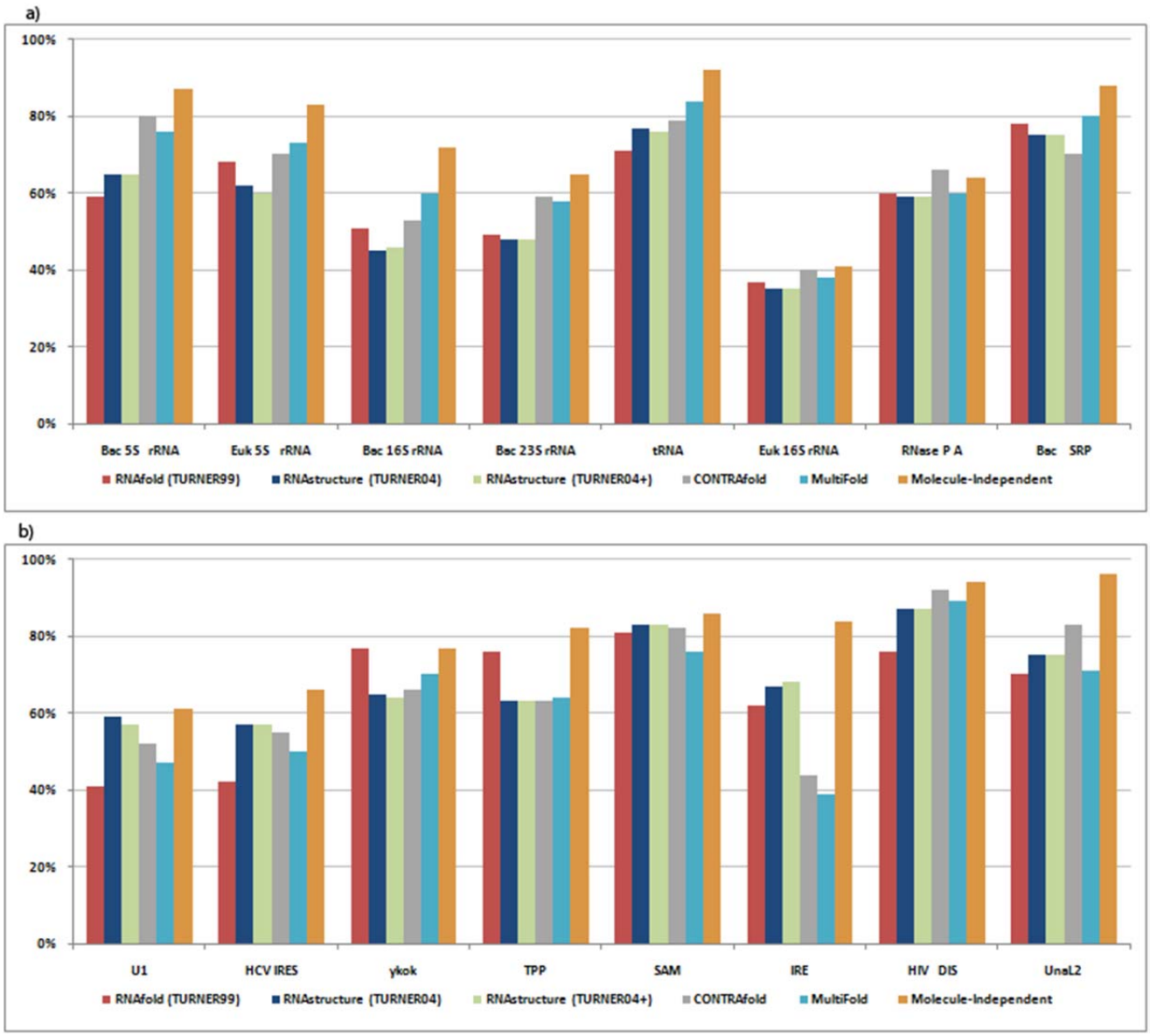

Figure S4: RNA secondary structure prediction accuracies for five RNA folding programs: RNAfold, RNAstructure (TURNER04 and (TURNER04+), ContraFold, MultiFold and RNAfold using statistical potentials. 80% of the sequences of each RNA molecular class were used for the training set of the statistical potentials. Testing results shown here are for the remaining 20%. Results for sixteen RNA molecular classes are divided into a) bacterial 5S rRNA, eukaryotic 5S rRNA, bacterial 16S rRNA, bacterial 23S rRNA, tRNA, eukaryotic 16S rRNA, RNase P A and bacterial SRP; b) U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader, TPP and SAM riboswitches, iron response element (IRE), HIV type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element.

The second method for testing our statistical potentials involved using nine control RNA molecular classes **(see methods)** that were not used in the generation of the statistical potentials and comparing our results against the results of the four other RNA folding programs. The hairpin statistical potentials had only small changes on most of molecular classes **(Fig. S5)**. Seven of the nine showed changes from 3% lower (GEMM) to no change relative to the original RNAfold. RNase P B (5% lower) and R2 (11% higher) experienced larger changes.

The results for internal loop statistical potentials showed slightly larger changes **(Fig. S5)**. Four of the nine showed no changes or slight decreases up to %3. HDV, RNase P B and archaeal 16S rRNA showed increases from 5% to 6%, but GEMM cis-regulatory element and R2 RNA element showed large decreases of 12% and 14%, respectively **(Supplemental Data, see Excel file Accuracies.xlsx)**.

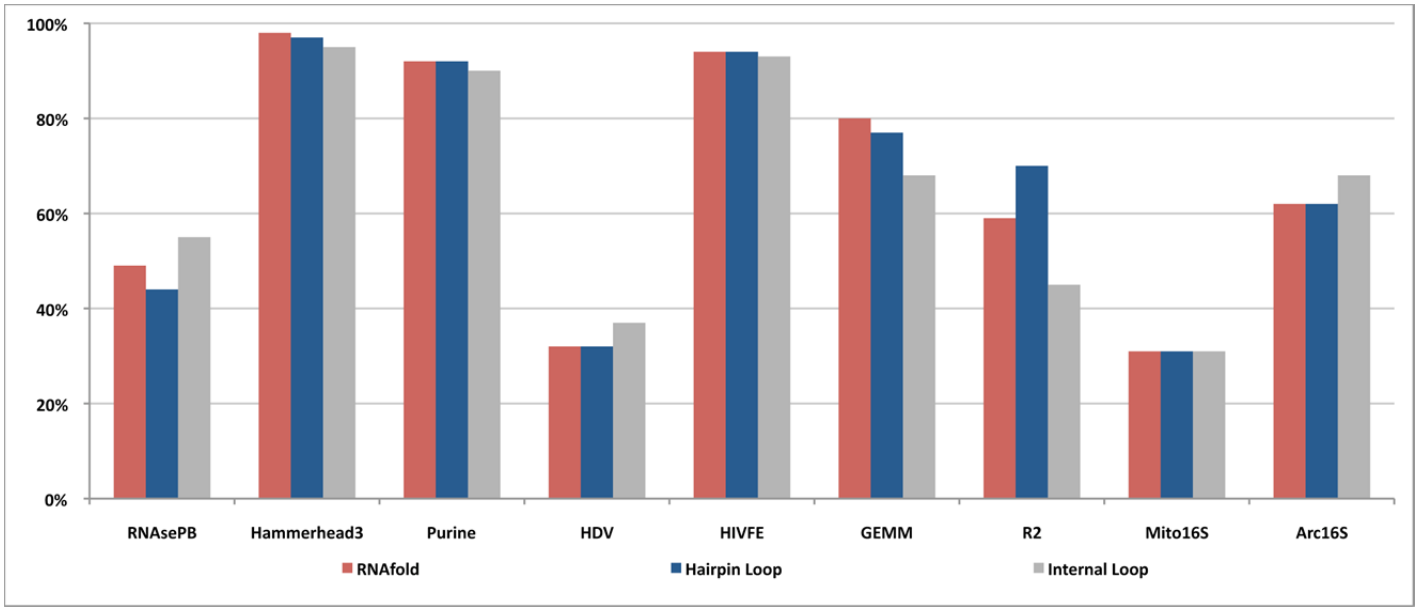

Figure S5: RNA secondary structure prediction accuracies for nine RNA molecular classes for RNAfold using 1) unmodified RNAfold, 2) molecule-independent hairpin loop statistical potentials; 3) molecule-independent internal loop statistical potentials. The experimentally derived energetic parameters for base pair stacks and multi-stem loops were used with both the hairpin internal loop statistical potentials. The results are for RNase P B, Hammerhead III ribozyme, purine riboswitch, hepatitis delta virus ribozyme (HDV), HIV ribosomal frameshift signal, GEMM cis-regulatory element, R2 RNA element and mitochondrial and archaeal 16S rRNA.

The next step for testing involved combining the hairpin and internal loop statistical potentials and testing them on the control sequences. On average, over these nine RNA molecular classes, our combined statistical potentials slightly outperformed RNAfold (TURNER99) by 2%, RNAstructure (TURNER04) by 3%, RNAstructure (TURNER04+) by 3% and CONTRAfold (.1%) **(Fig. S6)**. MultiFold scored 1% better than our statistical potentials. Our statistical potentials predicted the RNA secondary structure better in 26 out of 45 direct comparisons and equaled in two more comparisons **(Supplemental Data, see Excel file Accuracies.xlsx)**.

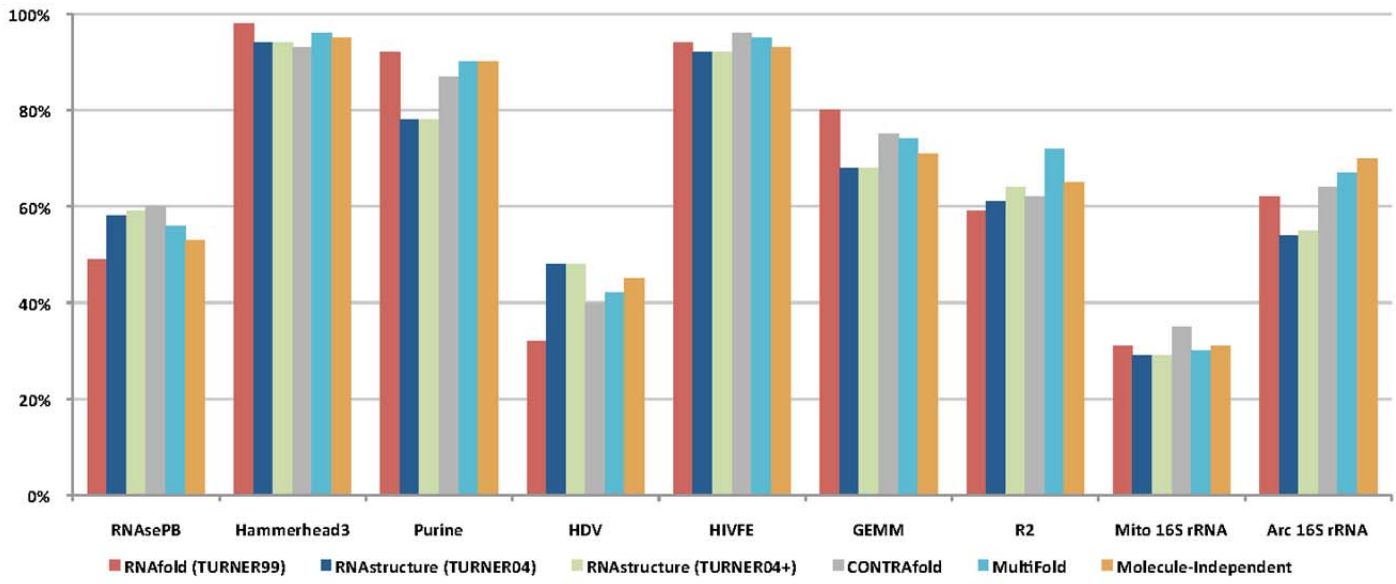

Figure S6: RNA secondary structure prediction accuracies for four RNA folding programs: RNAfold, RNAstructure (TURNER04 & TURNER04 plus newer tri- and tetraloop thermodynamic parameters), CONTRAfold, MultiFold and RNAfold using statistical potentials. Results for nine RNA molecular classes: RNase P B, Hammerhead III ribozyme, purine riboswitch, hepatitis delta virus ribozyme (HDV), HIV ribosomal frameshift signal, GEMM cis-regulatory element, R2 RNA element and mitochondrial and archaeal 16S rRNA.

The average standard deviations for the histogram in Figure S6 of the paper for each program for each of the nine control RNA molecular classes are shown as error bars in Figure S7. On average, our new statistical potentials had a similar standard deviation as RNAstructure and slightly higher for the other three programs **(Supplemental Data, see Excel file Accuracies.xlsx)**.

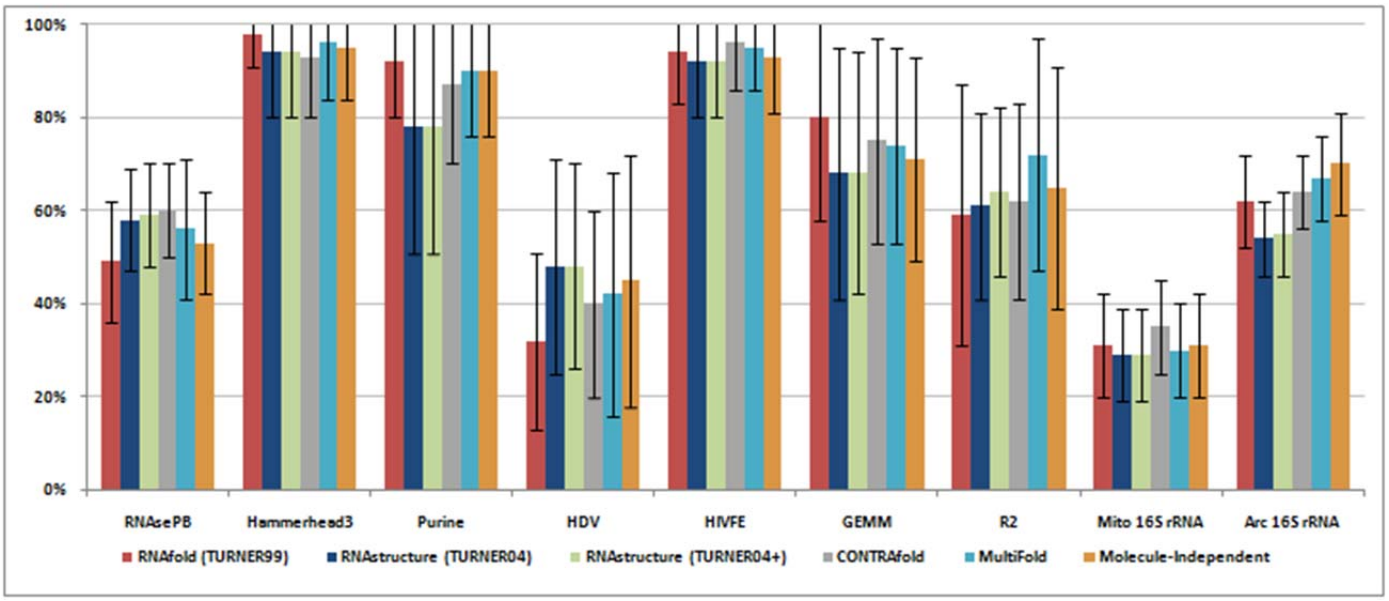

Figure S7: RNA secondary structure prediction accuracies plus standard deviations for five RNA folding programs: RNAfold, RNAstructure (TURNER04 and (TURNER04+), ContraFold, MultiFold and RNAfold using statistical potentials. The results are for RNase P B, Hammerhead III ribozyme, purine riboswitch, hepatitis delta virus ribozyme (HDV), HIV ribosomal frameshift signal, GEMM cis-regulatory element, R2 RNA element and mitochondrial and archaeal 16S rRNA.

Table S1: Number of sequences and their average length for eight RNA molecular classes used in determining prediction accuracy of RNA secondary structure folding programs.

	Bac 5S rRNA	Euk 5S rRNA	Bac 16S rRNA	Bac 23S rRNA	tRNA	Euk 16S rRNA	RNase P A	Bac SRP
# of Sequences	230	310	1062	64	2112	220	273	1049
Average Length	112	107	1463	2889	72	1677	336	100

Table S2: Number of sequences and their average length for eight RNA molecular classes used in determining prediction accuracy of RNA secondary structure folding programs.

	U1	HCV IRES	ykok	TPP	SAM	IRE	HIV DIS	UnaL2
# of Sequences	837	550	188	726	589	371	136	542
Average Length	159	214	168	102	104	29	40	53

Table S3: Number of sequences and their average length for four RNA molecular classes used in determining prediction accuracy of RNA secondary structure folding programs.

	RNase P B	Hammerhead 3	Purine	HDV
# of Sequences	114	84	133	33
Average Length	366	55	100	91

Table S4: Number of sequences and their average length for five RNA molecular classes used in determining prediction accuracy of RNA secondary structure folding programs.

	HIVFE	GEMM	R2	Mito 16S	Arc 16S
# of Sequences	145	162	15	128	143
Average Length	51	86	215	1053	1455

1. Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. & Murphy, K. P. (2010). Computational approaches for RNA energy parameter estimation. *RNA* **16**, 2304-18.