BJRNAFold: Prediction of RNA Secondary Structure Base on Constraint Parameters

Wuju Li and Xiaomin Ying
Center of Computational Biology, Beijing Institute of Basic Medical Sciences, PO Box 130(3), Beijing 100850, China
Email : wujuli@yahoo.com, liwj@nic.bmi.ac.cn

ABSTRACT: Predicting RNA secondary structure as accurately as possible is very important in functional analysis of RNA molecules. However, different prediction methods and related parameters including terminal GU pair of helices, minimum length of helices, and free energy systems often give different prediction results for the same RNA sequence. Then, which structure is more important than the others? i.e. which combinations of the methods and related parameters are the optimal? In order to investigate above problems, first, three prediction methods, namely, random stacking of helical regions (RS), helical regions distribution (HD), and Zuker's minimum free energy algorithm (ZMFE) were compared by taking 1139 tRNA sequences from Rfam database as the samples with different combinations of parameters. The optimal parameters are derived. Second, Zuker's dynamic programming method for prediction of RNA secondary structure was revised using the above optimal parameters and related software BJRNAFold was developed. Third, the effects of short-range interaction were studied. The results indicated that the prediction accuracy would be improved much if proper short-range factor were introduced. But the optimal short-range factor was difficult to determine. A user-adjustable parameter for short-range factor was introduced in BJRNAFold software.

1 INTRODUCTION

RNA molecules play an important role in many biological processes and knowing their structures is important in understanding their functions. In view of the difficulties in the experimental determination of RNA structures, the theoretical methods for RNA secondary structure prediction are often used. According to Higgs [1], there are four sorts of prediction methods, which are dynamic programming algorithms [2, 3], kinetic folding algorithms [4], genetic algorithms [5] and comparative methods [6, 7, 8]. When a set of phylogenetic-related RNA sequences are available, comparative methods can be used to find the conservative secondary structures. When there are only one or a few known sequences for an RNA, the first three sorts of methods can be used to predict RNA secondary structures. Therefore, molecular biologists have several choices to predict RNA secondary structures. But, when they face different predicted structures for the same RNA sequence, which structure is more rational? More generally, which prediction method is better? In addition, several RNA secondary structure prediction-related parameters can also affect the prediction results, which include terminal GU pair of helices, the minimum length of helices, and free energy systems [9, 10]. These parameters are the basic parameters in the field of prediction of RNA secondary structures. Then, which combinations of parameters are the optimal?

In this study, the above problems are investigated systematically. First, three prediction methods RS [11], HD [12], and ZMFE [3,13] are compared by taking 1139 tRNA sequences from Rfam database as the sample [14]. The optimal parameters and their combinations were found (See table 1 for detail information). For example, no matter whether terminal GU pair of helices is permitted or not, the best minimum length of helices for both RS and HD is 3 base pairs. Second, Zuker's dynamic programming method for RNA secondary structure prediction was revised using the above optimal parameters and related software BJRNAFold was developed. Compared to the RS method, BJRNAFold not only provides higher prediction accuracy for 1139 tRNA sequences, but also runs faster. It can be used to fold long RNA sequence ($\leq$1500 n.t. at present). In addition, both BJRNAFold and MZFE methods give the near same prediction accuracy for 9 RNA families with identity $\geq$ 72. Therefore, the studies here provide an alternative selection for molecular biologists to predict RNA secondary structures.

2 DATA AND METHODS

2.1 RNA sequences

There are 350 RNA families in the database [14]. The RNA families with the total number of seed sequences $>$ 50 were used for the evaluation of prediction methods. The number of such RNA families is 27. Because the sequence length is too long for RF00010, RF00177, and RF00023 families, it has some difficulty to predict RNA secondary structures using Mfold program in batch mode [13]. We finally selected the rest 24 families for evaluation. The detail information for these families was given in table 2. For tRNA family (RF00005), even through there are 1139 tRNA sequences, the phylogenetic analysis only provides 957 cloverleaf structures. Therefore, in cloverleaf structure based prediction accuracy calculation, we take 957 as the denominator.

2.2 Free energy systems and free energy calculation

Four free energy systems T3.0, T25, T37, and T42 were used. T3.0 (37^0C) is the latest free energy system [10], which was downloaded from Zuker's web page [13]. T25, T37, and T42 are correspondent to the temperature 25^0C, 37^0C and 42^0C respectively [9]. In this study, multibranch loops were assigned zero free energy (case I) or treated as interior loops (case II) [3]. In addition, if the loop size of hairpin, interior, or bulge $\geq$30, the free energy for these

loops was defined as $\triangle G^0(n)=\triangle G^0(30)+1.75RT\ln(n/30)$, in which n is the number of unpaired nucleotides in the loop, R is the gas constant (1.987cal.mol^{-1}.K^{-1}), T is the temperature in K (e.g. 298.15 K for 25^{0}C) [18].

2.3 Definition of prediction accuracy for RNA secondary structure

The prediction accuracy was defined as the ratio of the number of correctly predicted base pairs to the known base pairs from phylogenetic analysis. For tRNA sequences, we also define it as the ratio of the correctly predicted cloverleaf structures to the total number of cloverleaf structures (957) from phylogenetic analysis at the same time.

2.4 Prediction of RNA secondary structure based on random stacking of helical regions (RS)

The basic assumption of RS is that the direction of RNA folding is to lower the free energy of the forming structures, but the final structures may not reach the minimum free energy state. The structures are calculated using Monte Carlo simulations. By repeating the simulations, we can obtain many secondary structures for a given RNA sequence. Among those structures, some maybe appear several times, and some maybe appear many times. Here we take the structure with the highest occurrence frequency as the predicted structure. Another important parameter is the number of simulations (i.e. how many secondary structures should be calculated). According to our previous study [11, 12, 19], 100 simulations are enough for tRNA sequences. With the increase of sequence length, the highest frequencies of the structures will become very small. Therefore, RS is only applicable for short RNA sequences. But the validation of the mathematical model of high-level expression of foreign genes [20] based on the RS method demonstrated that the RS method is efficient and useful [21, 22].

2.5 Prediction of RNA secondary structure based on helical regions distribution (HD)

There are three steps to predict RNA secondary structure. First, the frequencies of helical regions are calculated from a set of secondary structures generated by RS, and the helical region with the highest frequency was added to the current structure. Second, all incompatible helices to the current structure are deleted from the helices list. Third, the above two steps are repeated until no helices in the list. Finally, we take the current structure as the predicted structure, which is composed of the helices that appear most frequently in the set of structures. When RNA sequence is too long (>200), the HD method will become very slow.

2.6 Zuker's minimum free energy method (ZMFE)

Zuker's method can find optimal and sub-optimal secondary structures for a given RNA sequence. The related program is Mfold [13]. In this study, we only consider the minimum free energy secondary structure predicted by Mfold.

2.7 Optimal parameters for RNA secondary structure prediction

In order to obtain the optimal parameters for prediction of RNA secondary structure, the secondary structures of 1139 tRNA sequences were predicted using three methods RS, HD, and ZMFE. Through the comparison of prediction results, the optimal parameters were found: terminal GU pair of helical regions was prohibited, the best minimum length of helical regions was 3 bp, and T25 was the best free energy system. These are the basis for the following revised dynamic programming methods. See results 1 for detail comparison information.

2.8 Revised dynamic programming methods for prediction of RNA secondary structure

Zuker's dynamic programming method was revised using the above optimal parameters. The basic steps for the minimum free energy calculation in revised algorithm are similar to those for the ZMFE method. Two matrix $V_{n\times n}$ and $W_{n\times n}$ are used, in which n is the given RNA sequence length. For any two bases i and j ($i<j$), $V(i,j)$ stores the minimum free energy value for the fragment [i,j] when i and j, i+1 and j-1, and i+2 and j-2 are paired. Here we only consider G-C, A-U, or G-U base pair. Furthermore, terminal G-U pair of helices is not allowed. $W(i,j)$ saves the minimum free energy information for the fragment [i,j] no matter whether i and j, i+1 and j-1, and i+2 and j-2 are base paired or not. At first, we assign initial values for the matrix V and W (e.g. 10000). Next, we consider the following two cases. The first case is j - i=8. The second is j - i>8.

For the first case (j - i=8), if i and j, i+1 and j-1, and i+2 and j-2 are paired, a potential hairpin loop structure will be formed. Based on free energy system, $V(i,j)$, $V(i+1,j-1)$ and $V(i+2,j-2)$ will be assigned corresponding free energy. $W(i,j)$, $W(i+1,j-1)$ and $W(i+2,j-2)$ will be given the same values. If one of i and j, i+1 and j-1, or i+2 and j-2 are not paired, $V(i,j)$, $V(i+1,j-1)$, $V(i+2,j-2)$, $W(i,j)$, $W(i+1,j-1)$ and $W(i+2,j-2)$ will keep the initial values.

For the second case (j-i>8), if i and j, i+1 and j-1, and i+2 and j-2 are base paired, $V(i,j)$ is calculated as follows

E_1=FE_Hairpin(i,j) (1)

(Hairpin structure, i+3 and j-3 is not allowed to pair)

E_2=FE_Helix(i,j) (2)

(Helix structure, i+3 and j-3 is allowed to pair)

E_3=FE_Interior(i,j,i',j') (3)

(Interior loop structure, i' and j', i'+1 and j'-1 and i'+2 and j'-2 are paired. i'-i>1 and j-j'>1)

E_4=FE_Buldge(i,j,i',j') (4)

(Buldge loop structure, i' and j', i'+1 and j'-1 and i'+2 and j'-2 are paired. i'-i=1 or j-j'=1, but not both)

E_5=FE_Multi(i,j) (5)

(Multibranch loop structure, more than one helical regions are extended from the fragment [i,j])

$V(i,j)=\min(E_1, E_2, E_3, E_4, E_5)$ (6)

W(i,j) will be calculated as follows

$E_6=V(i,j)$ (7)

$E_7=W(i,j-1)$ (8)

$E_8=W(i+1,j)$ (9)

$E_9=\min(W(i,m)+W(m+1,j)-\delta)$ $(i<m<j)$ (10)

(δ is the short-range factor)

$W(i,j)=\min(E_6, E_7, E_8, E_9)$ (11)

When j–i increases gradually from 8 to sequence length–1, $V(i,j)$ and $W(i,j)$ will be filled the minimum free energy for the fragment [i,j]. Finally, $W(1,n)$ stores the minimum free energy for the whole RNA sequence. The recursive procedures are used to search the minimum free energy secondary structure. In addition, in formula (10), a short-range factor δ is introduced. When δ is given a positive value, the sub-structure from E_9 will support the

short-range interaction. Otherwise, the long-range interaction will be supported. The effects of the δ will be given in the results.

2.9 Database construction and evaluation of RNA secondary structures

Two RNA secondary structure databases for each studied RNA families were constructed. First database is derived from phylogenetic analysis (PRNA2D). Second is created using prediction results from Mfold server (MRNA2D). All sequences for each family were submitted to Mfold server, and related minimum free energy secondary structures were used to construct the MRNA2D. Based on PRAN2D and MRNA2D, the prediction accuracies for Mfold were calculated. Finally, the prediction accuracies for each family were also calculated by comparing the prediction results from BJRNAFold to the corresponding secondary structures in PRNA2D database.

3 RESULTS

3.1 Comparison of three prediction methods RS, HD, and ZMFE

Based on the prediction results for 1139 tRNA sequences, three prediction methods RS, HD, and ZMFE were compared. During the prediction, no priori information was used. For RS method, the structures with the highest frequencies were taken as the predicted structures (in 100 simulations). For HD method, only one structure was predicted for each sequence. For ZMFE, only minimum free energy secondary structure was considered. From table 1, we can see clearly that the prediction accuracies are strongly related to the different prediction methods and related parameters. Furthermore, following conclusions were obtained. First, for both RS and HD, the prediction accuracies with terminal GU pair of helices prohibited are higher than the corresponding accuracies with GU pair of helices permitted. Second, for most combinations of parameters (i.e. each row in Table 1), the prediction accuracies from RS are higher than those from HD. Third, if terminal GU pair of helices is permitted, the best free energy system is T3.0; If terminal GU pair of helices is prohibited, the best free energy system is T25. Fourth, no matter whether terminal GU pair of helices is permitted or not, the best minimum length of helices is 3 bp for both RS and HD. Fifth, for the same method and parameter combinations, giving multibranch loops zero free energy will generate better prediction results than taking them as interior loops. Sixth, the optimal parameter combinations were found for both RS and HD. The related parameters are free energy system T25, no terminal GU pair of helices and the minimum length of helical regions 3 bp. The corresponding prediction accuracies are 79.46% and 78.29% of known base pairs or 54.65% (523/957) and 52.14% (499/957) of known cloverleaf structures respectively, which are far larger than 66.71% or 32.92% (315/957) from MZFE. Therefore, the RS outperforms HD and MZFE methods in some cases. But the following two shortcomings make the RS inappropriate for longer sequences. First, the frequencies of dominant structures will become very small with the sequence length > 120. Second, it is time consuming. In order to overcome the shortcomings of the RS method, Zuker's dynamic programming method for prediction of RNA secondary structure was revised using above optimized parameters and related program BJRNAFold was developed.

3.2 Prediction of tRNA secondary structures using BJRNAFold

The secondary structures of all 1139 tRNA sequences were predicted using BJRNAFold. The results indicated that the prediction accuracy was 76.13% of known base pairs or 41.80% (400/957) of known cloverleaf structures when the multibranch loops were given zero free energy. When the multibranch loops were treated as interior loops, the prediction accuracy was 82.01% of known base pairs or 56.53% (541/957) of known cloverleaf structures. Therefore, BJRNAFold outperforms RS method for tRNA sequences. Moreover, the running speed for BJRNAFold is much faster than RS method. BJRNAFold has the same time complexity as MZFE.

3.3 Prediction of RNA secondary structures for 24 Rfam members

The secondary structures for all sequences from 24 RNA families were predicted using MFold and BJRNAFold. The prediction accuracies for each family were calculated by comparing the prediction results to the secondary structures from phylogenetic analysis. The detail results were provided in Table 2. From table 2, we can obtain the following results. First, the average prediction accuracy for MFold is 73.06% for all 24 RNA families. For BJRNAFold, the average prediction accuracies 68.14% (case I) and 70.66% (case II) are less than 73.06% (Mfold). Second, for the RNA families with average identity ≥ 72 (RF00250, RF00175, RF00260, RF00229, RF00048, RF00026, RF00032, RF00002, and RF00008), the average prediction accuracy is 72.69% for MFold program. For BJRNAFold, the average prediction accuracies are 71.61% (case I) and 72.88% (case II) respectively. Both MFold and BJRNAFold provide near the same prediction accuracies.

3.4 Effect of short-range factor

Short-range factor was introduced in formula 10. Different values for short-range factor often lead to different prediction accuracies. For example, we let δ vary from 0.0 to 10.0 with the step size 0.1. For each δ value, the prediction accuracies for 1139 tRNA sequences were calculated using the software BJRNAFold. The related results were displayed in Figure 1. From Figure 1, we can see that the prediction accuracies were improved with the increase of δ. The maximum prediction accuracy 91.46% of known base pairs or 89.24% (854/957) of known cloverleaf structures was reached at $\delta = 3.7$ and 4.25 respectively, which was far larger than 79.46% of known base pairs or 54.65% (523/957) of known cloverleaf structures from the RS method. After $\delta = 4.25$, the prediction accuracies decreased with the increase of δ value. Therefore, with the introduction of proper δ value, BJRNAFold give better prediction accuracies. In fact, from table 2, we can see that the introduction of proper short-range factor will also improve prediction accuracies for other RNA families. But present difficult point is how to determine the optimal short-range factor. Therefore, we provide a user-adjustable parameter for short-range factor in BJRNAFold program.

4 DISCUSSIONS

In this study, we have finished the following works. First, based on 1139 tRNA sequences from Rfam database, three prediction methods RS, HD, and MZFE were compared. The optimal parameters and their combinations were found. Second, Zuker's dynamic programming method was revised using the above parameters. Third, an evaluation system for RNA secondary structure prediction methods based on phylogenetic-derived secondary structures from Rfam database was constructed. Using this system, we can evaluate the reliability of different prediction methods. In order to correctly predict the secondary structures of known RNA sequences, future studies will emphasize the following two points.

First is the exact estimation of short-range factor. From the prediction results, we find that the proper value of short-range factor would lead to the significant improvement of prediction accuracies. At present, we train the short-range factor for each Rfam family. For new RNA sequences, it is very difficult to find the optimal short-range factors. It needs further study.

Second is to evaluate the different prediction methods and related parameter combinations using the evaluation system. At present, many prediction methods are presented [13, 15, 16, 17]. When we face so many prediction methods, we cannot help ask which method and related parameters is the best. In this study, only RS, HD, and ZMFE were compared. We intend to collect more prediction methods and compare them. From this comparison, we can provide the molecular biologists with the best methods for prediction of RNA secondary structures.

5 ACKNOWLEDGEMENTS

The work is supported by grants #30270315 and #30470411 from the National Natural Science Foundation of China, and grant #5042021 from Beijing Natural Science Foundation.

6 REFERENCES

[1] P. G. Higgs. RNA secondary structure: physical and computational aspects. Quarterly Reviews of Biophysics, 33: 199-253, 2000.

[2] R. Nussinov, and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl Acad. Sci. USA, 77: 6309- 6313, 1980.

[3] M. Zuker, and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9: 133- 148, 1981.

[4] J. P. Arabhams, M. Van Den Berg, E. Van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. Nucleic Acids Res., 18: 3035-3044, 1990.

[5] J. H. Chen, S. Y. Le, and J. V. Maizel. Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Res., 28: 991-999, 2000.

[6] J. Parsch, J. M. Braveman, and W. Stephan. Comparative sequence analysis and patterns of covariation in RNA secondary structure. Genetics, 154: 909- 921, 2000.

[7] I. L. Hofacker. Vienna RNA secondary structure server. Nucl. Acids. Res., 33: 3429-3431, 2003.

[8] F. Tahi, M. Gouy, and M. Regnier. Automatic RNA secondary structure prediction with a comparative approach. Computers and Chemistry, 26: 521 – 530, 2002.

[9] D. H. Turner, N. Sugimoto, J. A. Jaeger, C. E. Longfellow, S. M. Freier, and R. Kierzek. Improved Parameters for Prediction of RNA Structure. Cold Spring Harbor Symp. Quant. Biol., 52: 123-133, 1987.

[10] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J. Mol. Biol., 288: 911-940, 1999.

[11] W. J. Li, and W. J. Wu. Prediction of RNA secondary structure based on random stacking of helical regions. Acta Biophys.Sinica, 12: 213-218, 1996.

[12] W. J. Li, and W. J. Wu. Prediction of RNA secondary structure based on helical regions distribution. Bioinformatics, 14: 700-706, 1998.

[13] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31: 3406-3415, 2003.

[14] J. S. Griffiths, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. Nucleic Acids Res., 31: 439-441, 2003.

[15] B. Knudsen, and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucl. Acids. Res., 31: 3423-3428, 2003.

[16] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. J. Mol. Biol., 319: 1059-1066, 2002.

[17] Y. Ding, C. Y. Chan, and C. E. Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. Nucl. Acids. Res., 32: W135-W141, 2004.

[18] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. Proc. Natl. Acad. Sci. USA, 86: 7706-7710, 1989.

[19] X. M. Ying, H. Luo, J. C. Luo, W. J. Li. RDfolder: a web server for prediction of RNA secondary structure. Nucleic Acids Res., 32: W150-W153, 2004.

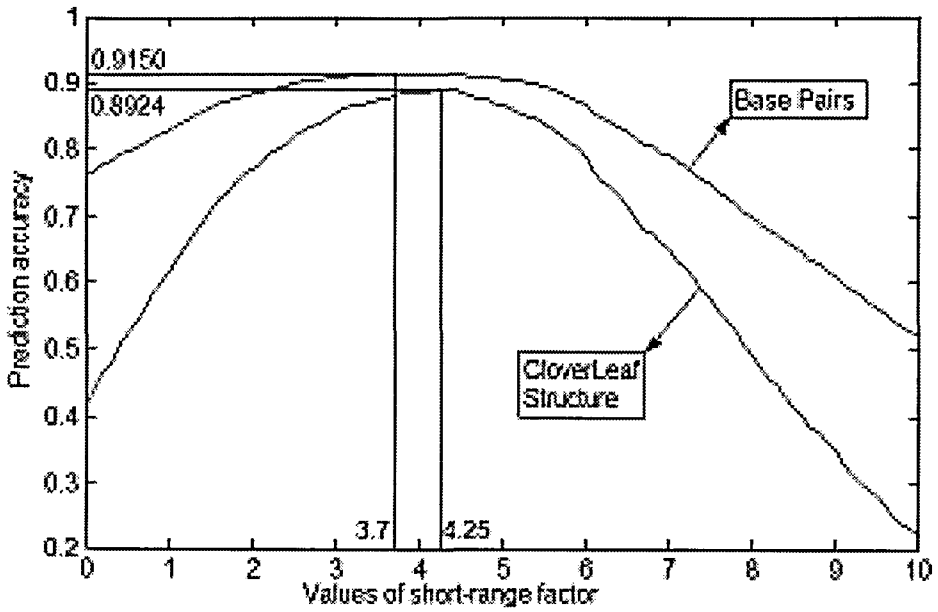
[20] W. J. Li, H. X. Lei, W. H. Pei, and J. J. Wu. GeneDn: for high-level expression design of heterologous genes in a prokaryotic system. Bioinformatics, 14: 884–885, 1998.

[21] W. H. Pei, B. F. Shen, and W. J. Li. Computer-aided design in high-expression of recombinant ricin-A chain in E.Coli. J. Cell. Mol. Immunol., 14: 33–36, 1998.

[22] W. H. Pei, M. R. Hu, W. J. Li, and B. F. Shen. The gene cloning and bioactivity of the expression product of the human FKBP12. Chin. J. Biochem. Mol. Biol., 16: 322–325, 2000.

7 FIGURES

Figure 1: Relationship between prediction accuracy and short-range factor



The relationship between the short-range factor and the prediction accuracy for 1139 tRNA sequences. The curve marked by "CloverLeaf Structure" stands for the percentage of predicted cloverleaf structures to the known cloverleaf structures (957). The curve marked by "Base Pairs" represents the percentage of predicted base pairs to the known base pairs from phylogenetic analysis. For cloverleaf structures, the maximum accuracy 89.24% (854/957) is reached at δ=4.25. For base pairs, the maximum accuracy 91.50% is reached at δ=3.70.

Table 1 - Prediction accuracies of three prediction methods RS, HD, and ZMFE for 1139 tRNA sequences.

Method	GU Pair	HL	FE	Accuracy*	N(R)*	Accuracy**	N(R)**	Method	GU Pair	HL	FE	Accuracy*	N(R)*	Accuracy**	N(R)**
RS	GU	2	T25	51.25	172(17.97)	43.5	91(9.51)	HD	GU	2	T25	52.48	169(17.66)	44.69	82(8.57)
			T3.0	64.51	277(28.94)	64.59	271(28.32)				T3.0	62.73	239(24.97)	62.48	233(24.35)
			T37	46.75	138(14.42)	38.44	61(6.37)				T37	48.77	138(14.42)	40.98	55(5.75)
			T42	45.46	152(15.88)	36.62	65(6.79)				T42	45.84	149(15.57)	37.88	52(5.43)
		3	T25	51.34	265(27.69)	42.33	114(11.91)			3	T25	50.45	248(25.91)	44.21	114(11.91)
			T3.0	65.27	336(35.11)	65.01	333(34.80)				T3.0	63.73	300(31.35)	63.61	307(32.08)
			T37	46.12	201(21.00)	38.01	99(10.34)				T37	46.24	182(19.02)	39.96	84(8.78)
			T42	44.26	200(20.90)	35.74	77(8.05)				T42	44.03	200(20.90)	36.12	64(6.69)
		4	T25	50.60	209(21.84)	42.63	75(7.84)			4	T25	49.9	197(20.59)	42.41	75(7.84)
			T3.0	64.69	299(31.24)	64.49	296(30.93)				T3.0	63.17	268(28.00)	63.05	270(28.21)
			T37	45.68	158(16.51)	38.35	68(7.11)				T37	44.37	136(14.21)	38.41	59(6.17)
			T42	43.39	144(15.05)	36.19	61(6.37)				T42	42.4	124(12.96)	36.07	51(5.33)
	NO GU	2	T25	79.35	467(48.80)	71.43	374(39.08)		NO GU	2	T25	78.04	451(47.13)	68.93	346(36.15)
			T3.0	68.76	345(36.05)	68.39	340(35.53)				T3.0	67.44	320(33.44)	67.29	314(32.81)
			T37	76.18	344(35.95)	66.20	247(25.81)				T37	74.13	320(33.44)	64.61	221(23.09)
			T42	74.04	318(33.23)	66.49	290(30.30)				T42	72.03	304(31.77)	63.94	259(27.06)
		3	T25	79.46	523(54.65)	71.63	432(45.14)			3	T25	78.29	499(52.14)	69.35	378(39.50)
			T3.0	0.23	378(39.50)	69.68	370(38.66)				T3.0	68.34	342(35.74)	69.18	349(36.47)
			T37	76.12	388(40.54)	67.06	305(31.87)				T37	74.23	370(38.66)	64.7	276(28.84)
			T42	73.53	349(36.47)	66.45	334(34.90)				T42	71.82	346(36.15)	63.43	291(30.41)
		4	T25	74.51	360(37.62)	69.31	330(34.48)			4	T25	73.63	362(37.83)	66.62	285(29.78)
			T3.0	69.01	309(32.29)	68.65	312(32.60)				T3.0	67.5	289(30.20)	67.55	292(30.51)
			T37	72.16	301(31.45)	66.45	285(29.78)				T37	70.58	298(31.14)	63.5	245(25.60)
			T42	70.44	268(28.00)	65.29	287(29.99)				T42	68.91	281(29.36)	62.51	244(25.50)
ZMFE	GU		T3.0	66.71	315(32.92)										

The secondary structures of 1139 tRNA sequences were predicted using the prediction methods RS, HD and ZMFE. For RS and HD methods, we considered all possible combinations of different parameters. 'GU' and 'NO GU' in GU Pair column stand for terminal GU pair of helices permitted and prohibited respectively. 2, 3, and 4 in HL column represent the different minimum length of helical regions. T25, T3.0, T37, and T42 in FE column stand for the different free energy systems (See data and methods for detail information). Accuracy column stand for the percentage of correctly predicted base pairs to the known base pairs from phylogenetic analysis. The integer number in N(R) column is the number of correctly predicted cloverleaf structures by the corresponding methods and parameters. The fraction in bracket in N(R) column is the percentage of predicted cloverleaf structures to the total number of cloverleaf structures from phylogenetic analysis (957). In addition, ZMFE stands for the prediction results from Mfold web server. '*' stands for the multibranch loops given zero free energy (case I). '**' stands for the multibranch loops treated as interior loops (case II).

Table 2. Prediction accuracy for 24 Rfam members

RF Code	Description	No.	Len.	Identity	TotalBp	Prediction accuracy									
						MFold		BJRNAFold*				BJRNAFold**			
								δ =0		Optimized δ		δ =0		Optimized δ	
						PBP	%	PBP	%	PBP	%	PBP	%	PBP	%
RF00032	Histone 3' UTR stem-loop	65	26	77	390	390	100.00	390	100.00	390	100.00	390	100.00	390	100.00
RF00260	HCV cis-acting replication element (CRE)	52	51	85	692	668	96.53	668	96.53	668	96.53	662	95.66	662	95.66
RF00008	Hammerhead ribozyme (type III)	84	55.4	72	1172	1121	95.65	1033	88.14	1092	0.93.17	1064	90.78	1093	93.26
RF00250	Trans-activation response element (TAR)	444	58.7	92	9767	9548	97.76	9164	93.83	9184	94.03	9093	93.10	9121	93.39
RF00048	Enterovirus cis-acting replication element	56	61	81	578	545	94.29	545	94.29	545	94.29	545	94.29	545	94.29
RF00031	Selenocysteine insertion sequence	65	64.4	43	820	648	79.02	579	70.61	597	72.80	583	71.10	596	72.68
RF00163	Hammerhead ribozyme (type I)	75	65.7	60	728	397	54.53	335	46.02	386	53.02	326	44.78	368	50.55
RF00005	tRNA	1139	73.3	43	21486	14334	66.71	16357	76.13	20152	93.79	17620	82.01	19933	92.77
RF00181	C/D box small nucleolar RNA 14q(I)/14q(II)	59	74.8	66	238	215	90.34	204	85.71	204	85.71	163	68.49	163	68.49
RF00029	Group II catalytic intron	116	85.7	55	1629	1490	91.47	1045	64.15	1297	79.62	1249	76.67	1309	80.36
RF00169	Bacterial signal recognition particle RNA	71	95	52	1302	1107	85.02	625	48.00	716	54.99	531	40.78	608	46.70
RF00026	U6 spliceosomal RNA	53	104.4	79	218	186	85.32	95	43.58	193	88.53	106	48.62	193	88.53
RF00059	TPP riboswitch (THI element)	141	104.6	52	2468	1566	63.45	1586	64.26	1791	72.57	1490	60.37	1693	68.60
RF00162	SAM riboswitch (S box leader)	71	110.5	66	1543	1183	76.67	1035	67.08	1242	80.49	1120	72.59	1260	81.66
RF00001	5S ribosomal RNA	602	116.6	59	9600	7276	75.79	5354	55.77	6309	65.72	5805	60.47	6572	68.46
RF00175	Retroviral Psi packaging element	173	117.4	87	4532	2826	62.36	2686	59.27	4302	94.92	2929	64.63	4309	95.08
RF00002	5.8S ribosomal RNA	63	152	77	671	450	67.06	396	59.02	559	83.31	417	62.15	569	84.80
RF00003	U1 spliceosomal RNA	54	158.8	62	868	497	57.26	374	43.09	604	69.59	465	53.57	598	68.89
RF00004	U2 spliceosomal RNA	77	184.8	60	2809	2294	81.67	1809	64.40	2198	78.25	1695	60.34	2178	77.54
RF00174	Cobalamin riboswitch	82	209.8	47	2130	1316	61.78	1027	48.22	1379	64.74	1100	51.64	1449	68.03
RF00230	T-box leader	67	243.9	43	986	188	19.07	181	18.36	441	44.73	201	20.39	416	42.19
RF00229	Picornavirus internal ribosome entry site (IRES)	208	251.3	83	9969	4611	46.25	5067	50.83	5620	56.37	5192	52.08	5810	58.28
RF00017	Eukaryotic type signal recognition particle RNA	71	295.3	46	3581	2859	79.84	1794	50.10	2222	62.05	1903	53.14	2303	64.31
RF00036	HIV Rev response element	65	337.7	61	6475	6128	94.64	5334	82.38	5643	87.15	5164	79.75	5415	83.63
Total base pair or average prediction accuracy					84137	62358	73.06	57683	68.14	67734	80.01	59813	70.66	67553	79.80

RF Code and Description are provided by Rfam database. No. stands for the total number of seed sequences in each RNA family. Len. stands for the average length of seed sequences from each RNA family, which is different from the average length provided by Rfam database. Identity represents the average identity (Average %id) given by Rfam. TotalBp is the total number of base pairs from phylogenetic analysis for each RNA family. MFold and BJRNAFold columns stand for the prediction accuracies calculated by MFold and BJRNAFold respectively. PBP stands for the number of correctly predicted base pairs. '*' stands for the multibranch loops assigned zero free energy. '**' represents the multibranch loops treated as the interior loops