Published online 1 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 4 ***e26***
doi:10.1093/nar/gkr1081

Statistical evaluation of improvement in RNA secondary structure prediction

Zhenjiang Xu[1], Anthony Almudevar[2] and David H. Mathews[1,2,*]

[1]Department of Biochemistry and Biophysics and Center for RNA Biology and [2]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA

Received December 24, 2010; Revised October 26, 2011; Accepted October 28, 2011

ABSTRACT

With discovery of diverse roles for RNA, its centrality in cellular functions has become increasingly apparent. A number of algorithms have been developed to predict RNA secondary structure. Their performance has been benchmarked by comparing structure predictions to reference secondary structures. Generally, algorithms are compared against each other and one is selected as best without statistical testing to determine whether the improvement is significant. In this work, it is demonstrated that the prediction accuracies of methods correlate with each other over sets of sequences. One possible reason for this correlation is that many algorithms use the same underlying principles. A set of benchmarks published previously for programs that predict a structure common to three or more sequences is statistically analyzed as an example to show that it can be rigorously evaluated using paired two-sample *t*-tests. Finally, a pipeline of statistical analyses is proposed to guide the choice of data set size and performance assessment for benchmarks of structure prediction. The pipeline is applied using 5S rRNA sequences as an example.

INTRODUCTION

There has been an explosion in our understanding of roles for RNA in cellular processes and gene expression in recent decades. RNA can form complex three dimensional structure either alone or with proteins to catalyze RNA splicing (1), catalyze peptide bond formation (2), guide protein localization (3), and tune gene regulation (4,5). Prediction of RNA secondary structure, the set of base pairing interactions between A-U, G-U and G-C, facilitates the development of hypotheses that connect structure to function. It also underlies a number of applications such as non-coding RNA detection (6–8), RNA tertiary structure prediction (9), siRNA design (10), miRNA target prediction (11) and structure design (12). There are many algorithms that have been developed to apply to specific situations with their own strengths and limitations.

The performance of a given structure prediction method is usually benchmarked on a set of RNA families by comparing predicted structures with the known secondary structures. Two statistical scores, sensitivity and positive predictive value (PPV), are commonly tabulated to determine the accuracy of the prediction methods (13). Sensitivity is the fraction of known pairs correctly predicted and PPV is the fraction of predicted pairs in the known structure. The two scores are sufficient to evaluate the accuracy. Another measure, the Matthews correlation coefficient (MCC) (14,15), is used as a single score that summarizes both sensitivity and PPV. It is defined as: $\mathrm{MCC} = (\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})/\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$, where TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative base pairs. It can be approximated by the geometric mean of sensitivity and PPV (15). Although MCC is a more succinct measure, by combining sensitivity and PPV it loses the information about capability and quality of a given method.

Traditionally, the averages of the statistical scores, either over families or sequences, are calculated to compare the accuracies of different folding algorithms. In this contribution, a statistical test is introduced to evaluate the performance of RNA folding methods against each other. Specifically, the prediction accuracies of some methods are shown to be correlated over sequences, and the paired two-sample *t*-test is proposed to compare accuracies of methods. Finally, precision and power analyses are suggested to determine the necessary dataset size for the benchmark to control the probabilities of hypothesis testing errors. An example is shown for programs that predict a consensus structure for multiple sequences, using all 5S rRNA sequences with reference structures available (16).

*To whom correspondence should be addressed. Tel: +1 585 275 1734; Fax: +1 585 275 6313; Email: david_mathews@urmc.rochester.edu

METHODS

Calculation of prediction accuracy

The prediction results of 12 RNA folding algorithms are taken from a previous study for this statistical analysis (16). That study reported a new algorithm for predicting an RNA secondary structure common to three or more sequences. When calculating sensitivity and PPV, a predicted base pair, $i-j$, counted as a correctly predicted pair if $i-j$, $(i+1)-j$, $(i-1)-j$, $i-(j+1)$ or $i-(j-1)$ pair is in the known structure (17). This is important because structures are compared to structures determined by comparative sequence analysis where there can be uncertainty in the exact pair match (18). Additionally, thermal noise can cause the actual structure to sample multiple pairs. For example, there is evidence in the thermodynamic studies of single RNA bulges that multiple conformational states exist (19,20).

Statistical analyses

All the following analyses were performed with the R statistical environment (21). Scripts are available for download from the Mathews lab website, http://rna.urmc.rochester.edu. All the figures were plotted with the Lattice package in R (22).

Correlation coefficient. To evaluate the independence between prediction results by two RNA folding algorithms, the Spearman rank correlation coefficient was calculated. It is a Pearson correlation coefficient between the ranks of two variables. It is calculated using the equation: $\rho_{x,y} = \sum(x_i-\bar{x})(y_i-\bar{y})/\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}$, where x_i and y_i are the ranks of average sensitivity or PPV predicted by the two algorithms on the ith group of sequences, and $\bar{x}$ and $\bar{y}$ are the means of the ranks. In case of tied observations, the arithmetic average of the rank numbers was used (23).

t-*test*. *t*-Tests were performed to calculate *P*-values in two ways. Assuming their prediction results were uncorrelated, an incorrect assumption, independent two-sample Welch's *t*-tests were used to compare the sensitivities or PPVs predicted by two algorithms with the null hypothesis that the means of the two samples are the same. Alternatively, paired two-sample *t*-tests were performed. Essentially the differences of individual sensitivities or PPVs from two algorithms were calculated to reduce the problem to a one-sample *t*-test with null hypothesis that the mean of the differences is zero (24).

Precision and power analysis. The probability of Type I error, denoted by α, is the probability of rejecting a true null hypothesis. The precision analysis relates the width of the confidence interval to the sample size by controlling α. A $(1-\alpha)$ percent confidence interval of a two-tailed paired *t*-test is defined as $\delta \pm t_{\alpha/2,(n-1)}S_d/\sqrt{n}$, where δ is the sample mean of the differences between sensitivities (or PPVs) of two algorithms, S_d is the sample standard deviation, n is the sample size and $t_{\alpha/2,(n-1)}$ is the critical value for the t distribution with $(n-1)$ degrees of freedom at the significance level of α. For an observed S_d, determining sample size n such that the confidence interval does not contain the point of zero, i.e. $\left|t_{\alpha/2,(n-1)}S_d/\sqrt{n}\right| \le \delta$, gives the estimated sample size to reject the null hypothesis with the probability of Type I error no greater than α (25).

The power of a statistical test is the probability of rejecting a false null hypothesis. It equals $(1-\beta)$, where β refers to the probability of the Type II error, namely, the probability of failing to reject a false null hypothesis. The sample size required for a given power can be prospectively estimated as $n = V_\mathrm{d}(Z_{1-\beta}+Z_{1-\alpha/2})^2/\delta^2$, where δ, V_d are the hypothetical mean and variance of the differences between two groups of scores, α and β are the probabilities of Types I and II errors, and $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are Z statistics of a standard normal distribution at the $(1-\beta)$th and $(1-\alpha/2)$th quantiles (25). Conversely, *a posteriori* β can be calculated after the statistical test from the equation to show the confidence to accept the null hypothesis with the given sample size and α. While precision analysis is often used to define the width of the confidence interval, *a priori* power analysis is used more to estimate a sample size to define how small a difference can be detected and with what degree of certainty.

Sequential test. Sequential testing is an alternative to fixed sample size testing (26,27). It is designed to reduce sample sizes without impacting power. Sample size is adaptively determined based on the accumulated data in a sequential test. It avoids the need for a single sample size estimate that would be larger than required for many cases. The methodology on the sequential procedure proposed in this article is in the Supplementary Material. Briefly, to prevent inflation of the Type I error rate introduced by sequential testing as opposed to testing with a fixed sample size, an adaptive rejection rule needs to be defined.

RESULTS

Many structure prediction methods employ similar energy models, evolve from the same underlying algorithm, or are extended from a previous version. It is possible that the prediction results of those related methods are correlated. In other words, two programs may make similar errors in structure prediction that would result in similar prediction accuracies on a number of sequences. The correlation was tested on a set of benchmark results published previously (16). This benchmark consists of 400 tRNA sequences (28) predicted by 12 methods, namely, Fold (29), Dynalign (30), Multilign (16), FoldalignM (31), mLocARNA (32), MASTR (33), Murlet (34), RAF (35), RNASampler (36), RNAshapes (37), StemLoc (38) and RNAalifold (39). Fold finds the lowest free energy structure for a single sequence. Dynalign finds the lowest free energy structure that is conserved for two unaligned sequences. RNAalifold folds multiple RNA sequences, prealigned with ClustalW2 in this benchmark. The remaining methods predict secondary structures of multiple sequences without requiring that the sequences be aligned ahead of time.

In the benchmark, 400 tRNA molecules were randomly selected and divided into groups of 5, 10 or 20 sequences. A single calculation was run over multiple sequences (except for Fold and Dynalign). For Dynalign, one sequence is chosen to be used with each of the other sequences in the group for structure prediction, and for this one sequence, the structure from last Dynalign calculation is used for scoring (16). The structures predicted in a single group are dependent on each other because the algorithms predict consensus structures. Thus, the sensitivities and PPVs for each sequence from the same calculation were averaged and this average was used as a single data point in the following statistical analyses. The sample size for the 5-, 10- and 20-sequence group predictions of 400 tRNA sequences are 80 (= 400/5), 40 (=400/10) and 20 (=400/20), respectively. By treating a single calculation as a data point instead of a single sequence as a data point, the statistical analysis is conservative in determining statistical significance because some power is lost in the calculation. This is specific to statistical tests of methods that use multiple sequences to find a consensus structure.

Correlation analysis

The most widely used correlation analysis is the Pearson product–moment correlation. It, however, is a parametric statistic designed to test a linear relationship between two variables normally distributed along an interval. For RNA structure prediction, the sensitivities or PPVs appear to be distributed far from the normal distribution (Supplementary Figure S1). Because of this, the Spearman rank correlation coefficients, which are more robust to outliers with extreme value, were computed (23). It is a non-parametric measure calculated with the ranks rather than the values of the observations. Figure 1 shows the matrices of correlation coefficients between sensitivities or PPVs predicted by any two algorithms over tRNA.

Multilign and Dynalign are correlated because Multilign is based on multiple Dynalign calculations. The correlation coefficient decreases from over 0.84 to ~0.65 when the number of sequences in one Multilign calculation increases from 5 to 20. This decreased correlation is expected because increasing the number of sequences in the Multilign prediction weakens the influence of individual Dynalign calculations. This is more obvious in the scatter plots in Figure 2. The horizontal distances of points from the diagonal line in the scatter plot of the 10 or 20 sequence predictions are longer than those of 5 sequence predictions, showing that the decreased correlation coefficients are at least partially due to the further prediction improvement on a few sequences by increasing the sequence number from 5 to 20.

Statistical significance

To test whether an algorithm has better performance than another, the means of their sensitivities and PPVs are traditionally compared. Comparing mean scores calculated from a sample of sequences, however, is usually not sufficient to infer the relationships between two statistical populations, i.e. the prediction accuracies of two folding algorithms on all possible RNA sequences.

In addition to reporting means, many studies additionally report standard deviations, but this is also not enough information to determine whether two methods have significantly different performance. To statistically evaluate the prediction accuracy of a method compared to another, a two-tailed *t*-test can be used to infer whether their accuracy means are significantly different. The test should be two-tailed because the better method is unknown ahead of time.

Incorrectly, Welch's *t*-test of two samples could be performed. Welch's *t*-test is similar to Student's *t*-test but is intended for use with two samples having possibly unequal variances. This test is performed on sensitivities and PPVs predicted by any two methods. The *P*-values are shown in Figure 3A. The *P*-value represents the risk of Type I error of rejecting the null hypothesis, that is the probability of claiming the two methods predict with different accuracies when actually their accuracies are the same.

Welch's *t*-test does not require equal variance, but like the typical (Student) *t*-test, it assumes the population from which the sample is drawn is normally distributed and the two populations are independent. Although the distribution of calculated sensitivities or PPVs is unlikely to be normal, a *t*-test is still justified by Central Limit theorem (CLT) with a large sample size (i.e. $n \geq 30$). The CLT says that $(\bar{x} - \mu)/(\sigma/\sqrt{n}) \rightarrow N(0,1)$ when the sample size n is large, where $\bar{x}$ is the sample mean, μ and σ are the population mean and population standard deviation. Replacing σ with the sample standard deviation s gives the *t*-distribution of $(n-1)$ degrees of freedom, $(\bar{x} - \mu)/(s/\sqrt{n}) \rightarrow t(n-1)$ (40). The assumption of independence, however, is clearly violated as explained above by the fact that the methods have correlated prediction accuracy (Figure 1). This correlation results in a systematic overestimation of the *P*-values by Welch's *t*-test in comparison with the paired test.

Given the observed correlation in prediction accuracy, a paired two-sample *t*-test is the appropriate statistical test. It calculates the differences of each observation between two methods and tests whether the mean of these differences is significantly different from zero. In the presence of significant correlation, paired *t*-tests have greater statistical power than unpaired *t*-tests by canceling the correlation when the two groups being compared occur in natural pairs. As shown in Figure 3, the *P*-values calculated from independent two-sample *t*-test (A) tend to be systematically larger than those from paired *t*-test (B). Specifically, the differences of PPVs and sensitivities between Multilign and Dynalign are significant when judged using paired *t*-tests and an α chosen to be 5%. The significances, however, are obscured by the incorrect application of independent *t*-tests (Figure 3A).

Powering the benchmarks

For those comparisons with *P*-values larger than 0.05, the benchmarks fail to demonstrate that there is a significant difference in prediction accuracies between the two

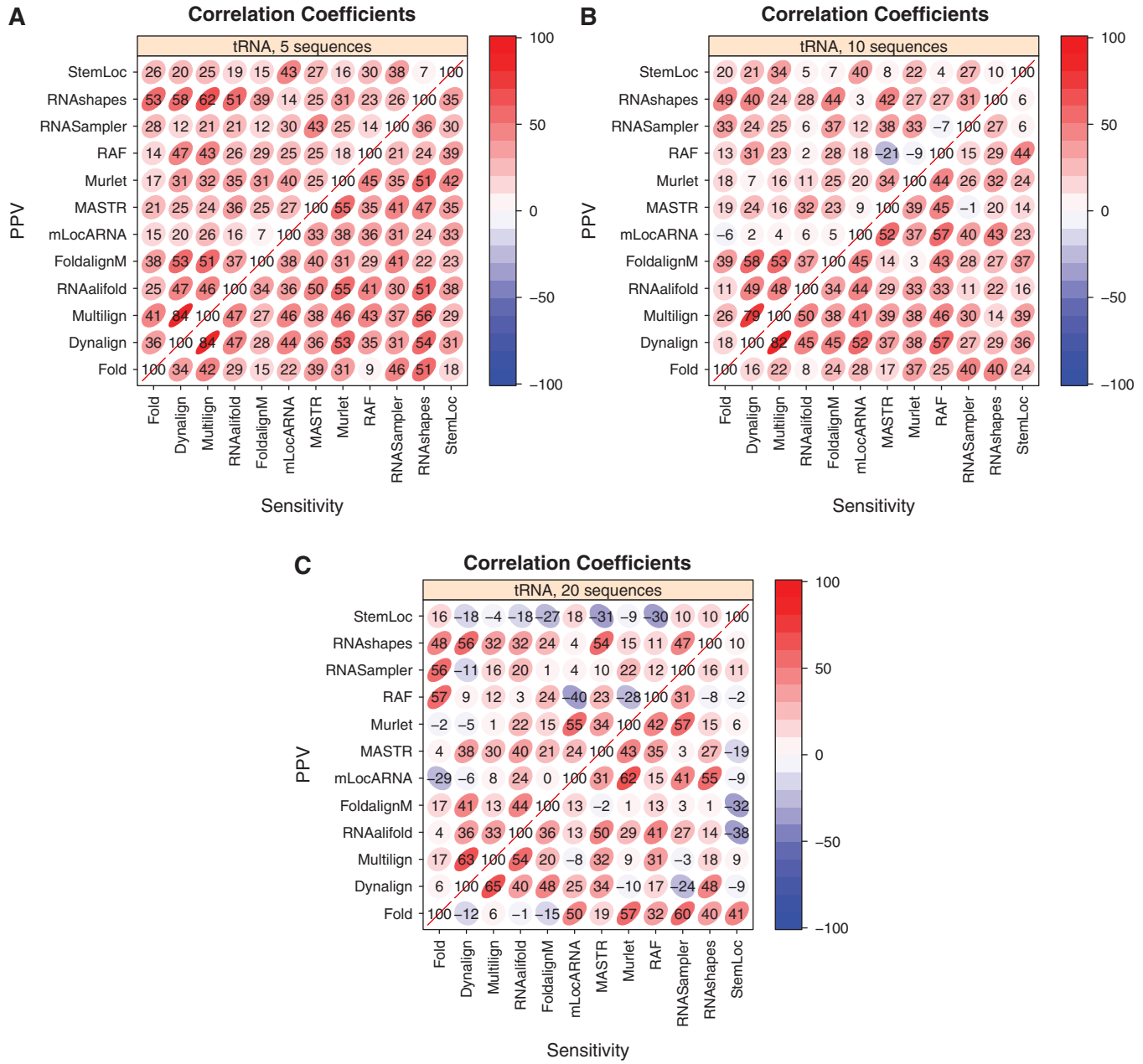

Figure 1. Spearman rank correlation coefficients between scores of RNA folding algorithms. The upper left triangle shows the correlation coefficients of PPV and the lower right triangle shows the correlation coefficients of sensitivity. The labeled numbers inside the ellipses indicate rounded correlation coefficients in percentage. The prediction results of tRNA in 5 (**A**), 10 (**B**) and 20 (**C**) sequence combinations are shown here. The colored ovals depict the extent of correlation (red) or anticorrelation (blue) between methods.

methods with a low Type I error. There is a subtle but important distinction between failing to demonstrate that there is difference and demonstrating there is no difference. The caveat is that the benchmark might lack the power to detect the difference due to the small sample size and large data variations, if the difference does exist. How big should the sample size be to detect small difference with controlled Type I error? This is non-trivial because benchmarks are costly in computer time, so choosing an optimally sized data set can save computation time when comparing two folding algorithms. On the other hand, if the null hypothesis cannot be rejected, Type II error, which is the probability of failing to reject a false null hypothesis, needs to be controlled. Conventionally the probability of Type II error is denoted by β and is chosen to be no larger than 0.2.

Based on that, a pipeline is proposed to statistically evaluate the performance of algorithms using pairwise comparisons (Figure 4). During the benchmark, data are accumulated sequentially over time until the available computation time or sequences run out or the null

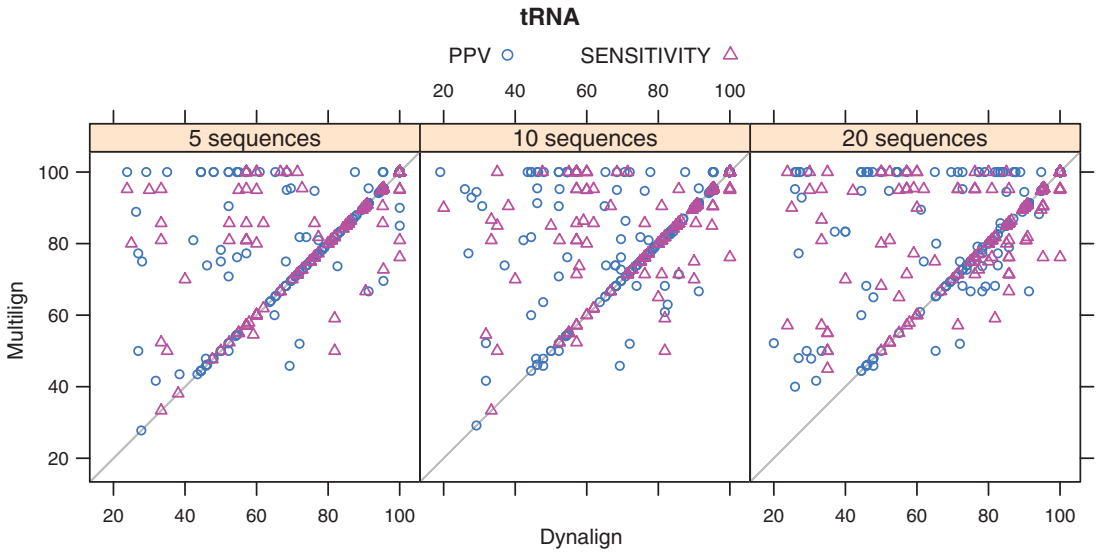

Figure 2. Scatter plot of Multilign scores versus Dynalign scores. A dot indicates the sensitivity (open triangle) or PPV (open circle) of a sequence predicted by Multilign (*y* axis) and Dynalign (*x* axis). The dots on the diagonal line mean the sensitivities or PPVs predicted by the two algorithms are the same. The sensitivities or PPVs of those sequences predicted by Multilign are better than those predicted by Dynalign if the dots are located above the diagonal line or worse if they are located below the diagonal line. The horizontal (upper triangle) or vertical (lower triangle) distances from the dots to the diagonal line are the differences of the scores between Multilign and Dynalign. Only the prediction results of tRNA using either 5 (left panel), 10 (middle panel) or 20 (right panel) sequences per Multilign calculations are shown here.

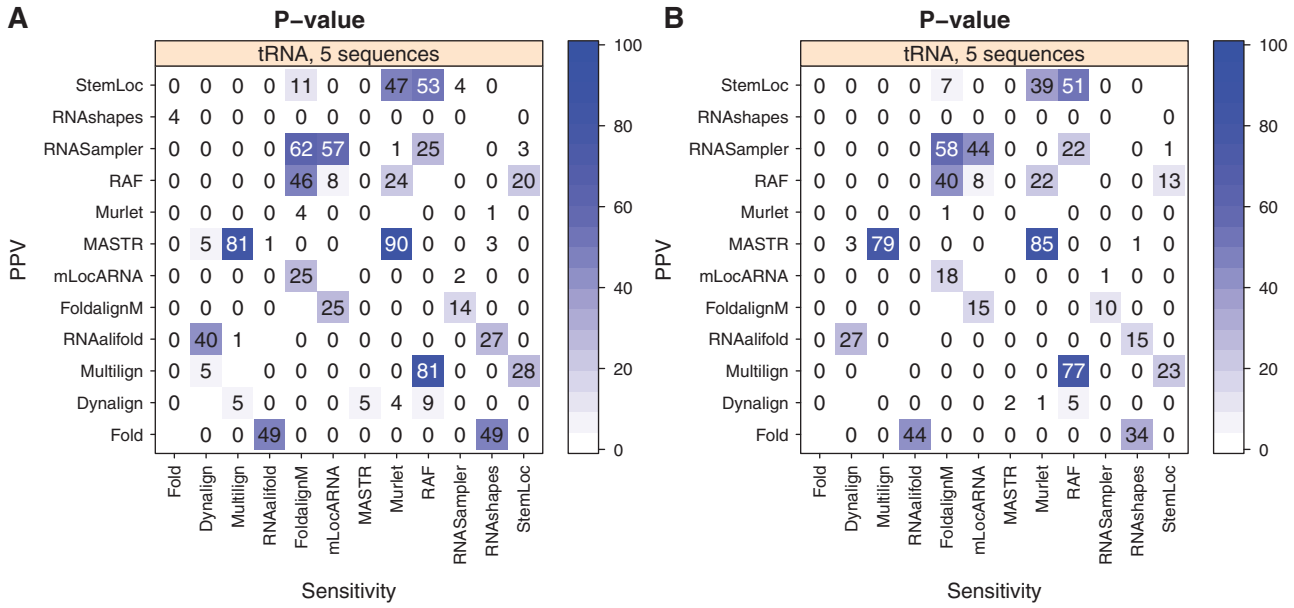

Figure 3. *P* values from *t*-tests for tRNA structure prediction. *P*-values in (**A**) are calculated with the independent two-sample Welch's *t*-test with null hypothesis that the average score predicted by the two algorithms are the same. *P*-values in (**B**) are calculated from the paired two-sample *t*-test. The upper left triangle shows the *P*-values of PPV and the lower right triangle shows the *P*-values of sensitivity. Values are expressed as rounded percentages. For the multiple (>2) sequence methods, the results of using five sequences in a single calculation are shown here.

hypothesis is proven or disproven. Thus a sequential paired two-sample *t*-test can be employed to reach an early conclusion regarding the comparison of prediction accuracy. This sequential style of interim estimate, however, can inflate the Type I error probability (41,42). Here the critical value is adjusted by simulations to accommodate this inflation. If the null hypothesis cannot be rejected at any of the interim tests, the power is calculated to see whether the null hypothesis can be accepted with high confidence. If no conclusion can be made, the process proceeds to the next stage of testing with benchmarks on an additional group of sequences. This step can be iterated until a conclusion is reached, computer resources are depleted, or until no more sequences are available.

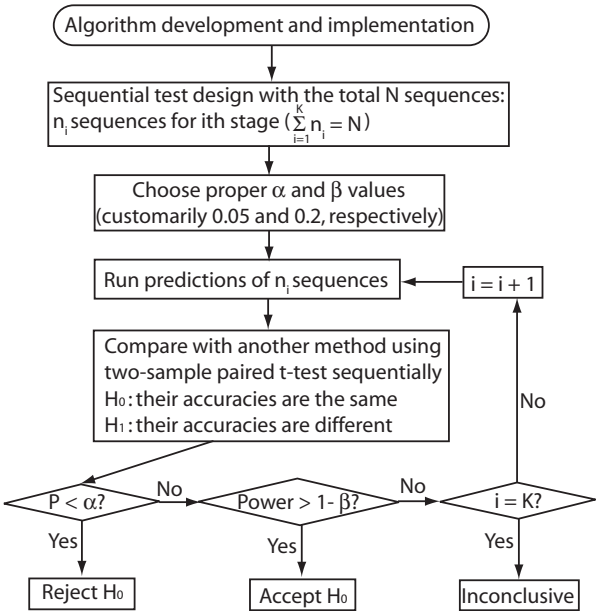

Figure 4. Statistical analysis pipeline. After the implementation of a new algorithm, its performance can be compared to another method with a sequential paired two-sample *t*-test. The total available sequences (N) are benchmarked in a number of stages (maximum of K). The next stage is needed only if no conclusion can be made with controlled Type I and II error probabilities. Note that α is adjusted as explained in the Methods section to prevent inflation of Type I error.

To illustrate the pipeline, the performances of the same 12 methods as above were evaluated on groups of five 5S rRNA sequences (43). In the Multilign paper, the sample size for 5S rRNA is only 20 (= 100 sequences/5 sequences per calculation). It is expanded with an additional 1095 sequences to reach the sample size of 239. Then the statistical analysis proceeded sequentially for up to 12 stages, with 20 more predictions done on an extra 100 sequences at each stage (except that the last one only has 95 sequences because all available sequences were used). At the beginning, a *t*-statistic for the first 20 predictions is calculated and compared to the adjusted critical values in the same manner as the paired two-sample *t*-test. If the null hypothesis is neither rejected nor accepted with defined Types I and II error probabilities, the *t*-test is repeated with 20 additional calculations. Many pairwise comparisons are shown to already have significant difference in the first stage (Figure 5). Take Multilign versus RNAalifold for example, Multilign is significantly better than RNAalifold for both sensitivity and PPV with the original 100 sequences. For some cases, the iterative sampling method does refine the test to be able to draw conclusions. For example, using the first 20 groups of sequences, the *a priori* power analysis estimates that 142 groups of sequences are needed to resolve the hypothesis of comparing the sensitivity between Multilign and Dynalign. This is much more than the 80 groups eventually used in the sequential test, which demonstrates that Multilign does have a higher average PPV than Dynalign. There are, however, also many inconclusive comparisons. For example, the null hypothesis that MASTR is as accurate as RNAalifold cannot be rejected, and it cannot be accepted either, for the inability to reject the null hypothesis may be due to the small power (0.18 and 0.20 for sensitivity and PPV), not because they truly have the same performance. Unfortunately, the available 5S rRNA sequences run out before sufficient power can be achieved to make a conclusion.

DISCUSSION

Because of the natural variability between measurements and the limitations on sampling, one cannot conclude that a method is better than another merely on the basis of means, especially when the difference in means is small. In this article, a statistical analysis is introduced to test the significance of improved RNA secondary structure prediction accuracy. It shows that the performance of algorithms can be correlated and a paired two-sample *t*-test can be used to compare the performance of two algorithms in terms of sensitivity or PPV. Even if there is no relationship between two RNA folding methods, a paired *t*-test is still justified because both methods are tested on the same set of sequences and thus their prediction scores are matching pairs. Another caveat is that a correlation coefficient of zero does not necessarily mean the two sets are uncorrelated. An example is the parabola function $y = x^2$, for which both the Spearman rank correlation coefficient and Pearson product–moment correlation coefficient are zero for a sample centered on $x = 0$, although there is perfect quadratic relationship between x and y.

Finally, a pipeline of evaluation is provided as illustrated in Figure 4. After the development and implementation of a new algorithm, its performance can be compared to another algorithm with a sequential procedure illustrated in the flow chart. If no conclusion can be made in the *i*th stage, the procedure moves on to the next stage with more sequences added into the benchmark. The sequential test used here has the advantages of reducing the total number of sequences required for the statistical test. In reality, with the broad range of the scores, the small differences between the average performances, and limited number of sequences available, it is often the case that which of two methods is better is statistically unknown. Because, however, the performance difference is usually small, the two methods may perform similarly in a practical sense on real data whether they are statistically different or not.

Although all the analyses above were performed on sensitivity and PPV, they essentially apply to other scores, such as MCC or the alignment score (sum of pairs score) without any modifications.

The method presented here, in Figure 4, provides the means for assessing whether the structure prediction performance of a program is significantly better than another on a benchmark set of sequences. It does not, however, determine whether one program is actually better than another. One first reason is that the benchmark set of sequences needs to be well chosen to represent the actual way the programs will be applied. A second

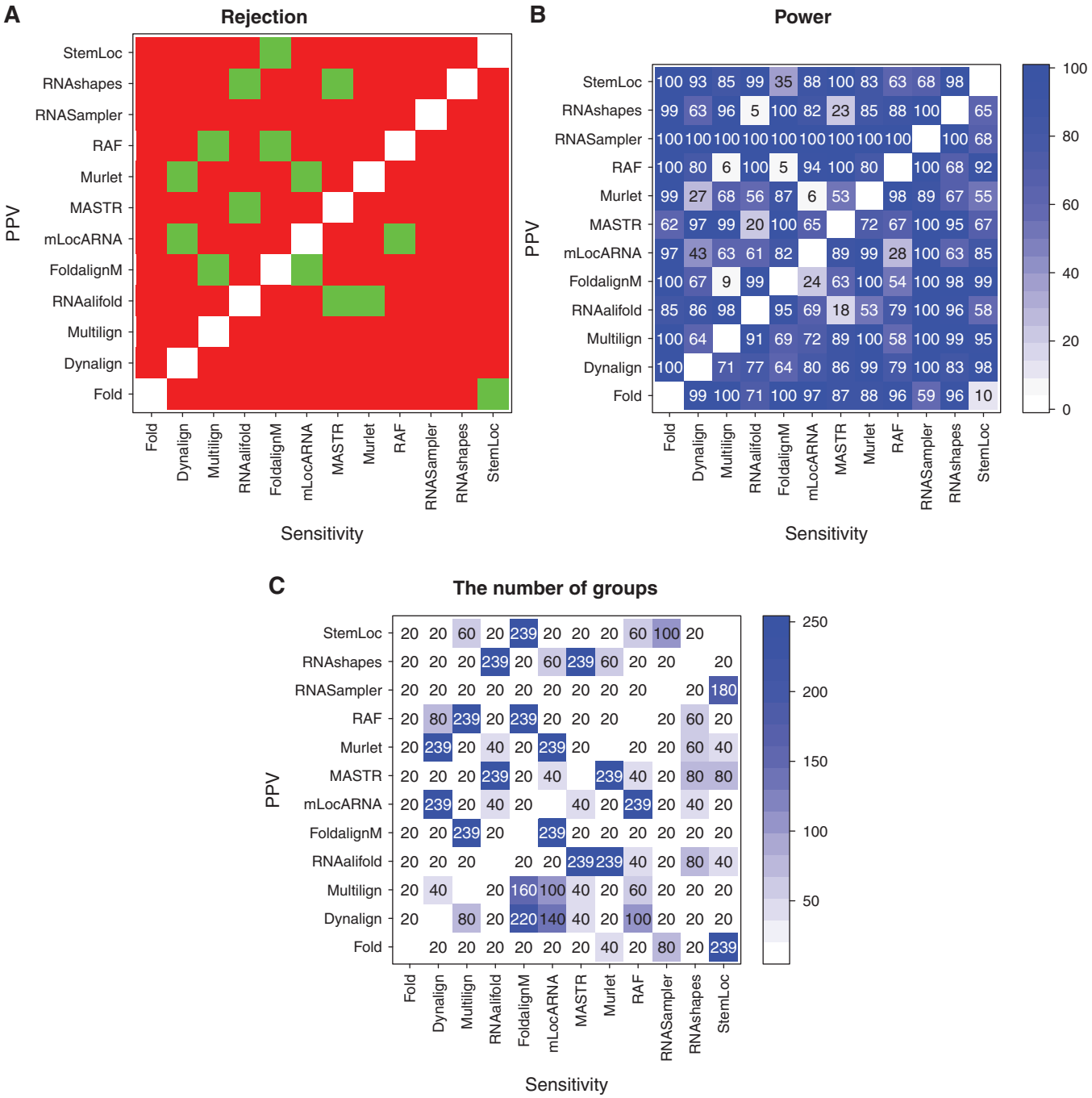

Figure 5. Example of 5S rRNA illustrating the flowchart of Figure 4. The 12 methods evaluated against each other on groups of five 5S rRNA sequences. Starting with 20 calculations (100 sequences), an additional group is benchmarked and statistically tested until the P value is smaller than α, power is greater than $1-\beta$, or the total 239 groups of sequences ran out. (**A**) The final conclusions. Red: null hypothesis rejected; green: not rejected. (**B**) The final powers between any two methods. These are expressed as rounded percentages. If the null hypothesis is not rejected but the power is larger than 0.8, then the null hypothesis can be accepted; otherwise, it is inconclusive. (**C**) The number of groups needed to test the hypothesis or the total available groups for those inconclusive comparisons (239 maximum groups of five sequences). In each panel, the upper triangle applies to PPV and the lower triangle to sensitivity. For reference, the average sensitivity and PPV for each program is provided in Supplementary Table S2.

reason is that programs may perform better on some families of sequences than on others, so a statistically better performance on one family of sequences may not result in better performance on a different family of sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Tables 1 and 2, Supplementary Figure 1, and Supplementary methods.

ACKNOWLEDGEMENTS

The authors thank Holger Hoos, University of British Columbia, and Charles Lawrence, Brown University, for insightful comments.

FUNDING

National Institutes of Health (grant R01HG004002 to D.H.M.). Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
2. Nissen,P., Hansen,J., Ban,N., Moore,P.B. and Steitz,T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
3. Jambhekar,A. and DeRisi,J.L. (2007) Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA*, **13**, 625–642.
4. Yanofsky,C. (1981) Attenuation in the control of expression of bacterial operons. *Nature*, **289**, 751–758.
5. Winkler,W.C., Nahvi,A., Roth,A., Collins,J.A. and Breaker,R.R. (2004) Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, **428**, 281–286.
6. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
7. Uzilov,A.V., Keegan,J.M. and Mathews,D.H. (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinform.*, **7**, 173.
8. Torarinsson,E., Sawera,M., Havgaard,J.H., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
9. Shapiro,B.A., Yingling,Y.G., Kasprzak,W. and Bindewald,E. (2007) Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.*, **17**, 157–165.
10. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
11. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.
12. Zadeh,J.N., Wolfe,B.R. and Pierce,N.A. (2011) Nucleic acid sequence design via efficient ensemble defect optimization. *J. Comput. Chem.*, **32**, 439–452.
13. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
14. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
15. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
16. Xu,Z. and Mathews,D.H. (2011) Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics*, **27**, 626–632.
17. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
18. Pace,N.R., Thomas,B.C. and Woese,C.R. (1999) Probing RNA structure, function, and history by comparative analysis. In: Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*, 2nd edn. Cold Spring Harbor Laboratory Press, New York, pp. 113–141.
19. Znosko,B.M., Silvestri,S.B., Volkman,H., Boswell,B. and Serra,M.J. (2002) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*, **41**, 10406–10417.
20. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
21. R Development Core Team. (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
22. Sarkar,D. (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York.
23. Glantz,S. (2005) *Primer of Biostatistics*, 6th edn. McGraw-Hill Medical, New York.
24. Dalgaard,P. (2008) *Introductory Statistics with R*, 2nd edn. Springer, New York.
25. Gerstman,B.B. (2009) *Basic Biostatistics: Statistics for Public Health Practice*, 1st edn. Jones and Bartlett Publishers, Sudbury, MA.
26. Siegmund,D. (1985) *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.
27. Wald,A. (1947) *Sequential Analysis*. Wiley, New York.
28. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
29. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.*, **11**, 129.
30. Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
31. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926.
32. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
33. Lindgreen,S., Gardner,P.P. and Krogh,A. (2007) MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, **23**, 3304–3311.
34. Kiryu,H., Tabei,Y., Kin,T. and Asai,K. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.
35. Do,C.B., Foo,C.-S. and Batzoglou,S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, i68–i76.
36. Xu,X., Ji,Y. and Stormo,G.D. (2007) RNA sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, **23**, 1883–1891.
37. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
38. Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinform.*, **6**, 73.
39. Hofacker,I.L. (2007) RNA consensus structure prediction with RNAalifold. *Methods Mol. Biol.*, **395**, 527–544.
40. Gosling,J. (1995) *Introductory Statistics*. Pascal Press, Sydney, Australia.
41. Shao,J. and Feng,H. (2007) Group sequential t-test for clinical trials with small sample sizes across stages. *Contemp. Clin. Trials*, **28**, 563–571.
42. Cui,L., Hung,H.M.J. and Wang,S.-J. (1999) Modification of sample size in group sequential clinical trials. *Biometrics*, **55**, 853–857.
43. Szymanski,M., Barciszewska,M.Z., Erdmann,V.A. and Barciszewski,J. (2002) 5S ribosomal RNA database. *Nucleic Acids Res.*, **30**, 176–178.