RNA (1998), *4*:669–679. Cambridge University Press. Printed in the USA.
Copyright © 1998 RNA Society.

Using reliability information to annotate RNA secondary structures

MICHAEL ZUKER[1] and ANN B. JACOBSON[2]

[1]Institute for Biomedical Computing, Washington University, St. Louis, Missouri 63110, USA
[2]Department of Microbiology, School of Medicine, State University of New York, Stony Brook, New York 11794, USA

ABSTRACT

A number of heuristic descriptors have been developed previously in conjunction with the *mfold* package that describe the propensity of individual bases to participate in base pairs and whether or not a predicted helix is "well-determined." They were developed for the "energy dot plot" output of *mfold*. Two descriptors, P-num and H-num, are used to measure the level of promiscuity in the association of any given nucleotide or helix with alternative complementary pairs. The third descriptor, S-num, measures the propensity of bases to be single-stranded. In the current work, we describe a series of programs that were developed in order to annotate individual structures with "well-definedness" information. We use color annotation to present the information. The programs can annotate PostScript files that are created by the *mfold* package or the PostScript secondary structure plots produced by the Weiser and Noller program XRNA (Weiser B, Noller HF, 1995, *XRNA: Auto-interactive program for modeling RNA*, The Center for Molecular Biology of RNA, Santa Cruz, California: University of California; Internet: ftp://fangio.ucsc.edu/pub/XRNA). In addition, these programs can annotate *ss* files that serve as input to XRNA. The annotation package can also handle structure comparison with a reference structure. This feature can be used to compare predicted structure with a phylogenetically deduced model, to compare two different predicted foldings, and to identify conformational changes that are predicted between wild-type and mutant RNAs.

We provide several examples of application. Predicted structures of two RNase P RNAs were colored with P-num information and further annotated with comparative information. The comparative model of a 16S rRNA was annotated with P-num information from *mfold* and with base pair probabilities obtained from the Vienna RNA folding package. Further annotation adds comparisons with the optimal foldings obtained from *mfold* and the Vienna package, respectively. The results of all of these analyses are discussed in the context of the reliability of structure prediction.

Keywords: base pair uncertainty; boxplot; energy dot plot; *mfold*; RNA folding; RNA secondary structure

INTRODUCTION

The development of suboptimal algorithms for RNA secondary structure prediction (Williams & Tinoco, 1986; Zuker, 1989a, 1989b) helped to mitigate the uncertainty of predicting a secondary structure of a single RNA sequence from thermodynamic data. The *mfold* algorithm (Zuker, 1989a, 1994) predicts suboptimal foldings as well as an "energy dot plot," which is a dot plot showing all possible base pairs that can participate in foldings within a specified increment of the predicted minimum folding energy. The collection of suboptimal folding predictions and, in the case of the Zuker algorithm, the energy dot plot, combine to give the user an idea of how well-determined a given prediction is.

A number of heuristic descriptors have been developed in conjunction with the *mfold* package that describe the propensity of individual bases to participate in base pairs and whether or not a predicted helix is "well-determined." These descriptors are P-num, S-num, and H-num. The first two were introduced early (Jaeger et al., 1989, 1990) and are computed for individual bases.

P-num is defined from the energy dot plot, and therefore depends on an (arbitrary) energy increment and whether or not the dot plot has been filtered to eliminate isolated base pairs or short helices. For the ith base in a molecule with n bases, $P\text{-}num(i)$ is the total number of dots in the ith row and column of the dot plot. In simple words, $P\text{-}num(i)$ is the total number of different base pairs that can be formed using the ith base in all foldings within the prescribed energy increment. If $P\text{-}num(i)$ is large, and this is a relative term,

Reprint requests to: Michael Zuker, Institute for Biomedical Computing, Washington University, 700 S. Euclid Avenue, St. Louis, Missouri 63110, USA; e-mail: zuker@ibc.wustl.edu.

then the *i*th base is promiscuous in its association with other bases. We say that it is "poorly determined." In an ensemble of foldings, it will be single-stranded or paired with many different bases. In a particular folding, we cannot say with any certainty how this base will pair. If *P-num*(*i*) is 0, then the *i*th base must be single-stranded. Otherwise, P-num gives no information of the propensity to be single-stranded. This is furnished by S-num, defined next.

S-num is defined from a collection of optimal and suboptimal foldings, and is thus independent from any dot plot computations. In a group of *m* foldings, *S-num*(*i*) is the number of foldings in which base *i* is single-stranded, divided by *m*. Thus, *S-num*(*i*) is a sample probability that the *i*th base is single-stranded. A value of S-num that is close to 0 or 1 is "good" in the sense that it tells us with a high degree of confidence whether the base is paired or not paired, respectively. We say that a base is "well-determined" if S-num is near 1 (almost certainly single-stranded) or if S-num is near 0 (almost certainly base paired) and P-num is low.

At a later date (Zuker & Jacobson, 1995), we introduced the notion of H-num, which is an extension of P-num to helices. For a singe base pair, *i·j*, we define *H-num*(*i·j*) to be *P-num*(*i*) + *P-num*(*j*) − 1. This is simply the total number of base pairs that can be formed using the *i*th or *j*th bases, in all foldings within the chosen energy increment. For a helix, H-num is the average of these values for all the base pairs in the helix. Helices with relatively low H-num values are said to be "well-determined" and those with relatively high values are said to be "poorly determined." We used the H-num measure to demonstrate that "well-determined" helices in optimal foldings are more likely to be correct than "poorly determined" ones (Zuker & Jacobson, 1995). It is worth adding here that a "well-determined" helix does not have to be in an optimal folding.

The recursive computation of rigorous partition functions for the RNA secondary structure model (McCaskill, 1990) lead to the development of rigorous statistics to describe uncertainties in RNA folding predictions. The original work computes base pair probabilities and, as a direct consequence, probabilities that any base will be single- or double-stranded. Base pair probabilities are plotted in what is called a "boxplot." This is similar to the *mfold* energy dot plot, except that base pairs are plotted as black squares whose areas are proportional to the probability of that base pair. There is a probability cutoff, usually 10^{-6}, below which base pairs are not plotted. These ideas have been taken up and expanded on by a theoretical chemistry group at the University of Vienna. The resulting software has become to be known as the "Vienna (RNA) package" (Hofacker et al., 1994).

Although an experienced user of the *mfold* package can extract information relatively easily from the dot plot superposition of optimal and suboptimal foldings, we find that a color annotation of individual foldings simplifies the interpretation of the results that are obtained from these plots. The major innovation described in this work is the annotation of foldings with "well-definedness" information. The latter can be the P-num, S-num, or H-num measures computed from the current *mfold* package, or base pair probabilities and other measures as computed by the Vienna package.

Our annotation method was developed to annotate predicted secondary structures. However, structural models that have been created from comparative sequence analysis can also be annotated.

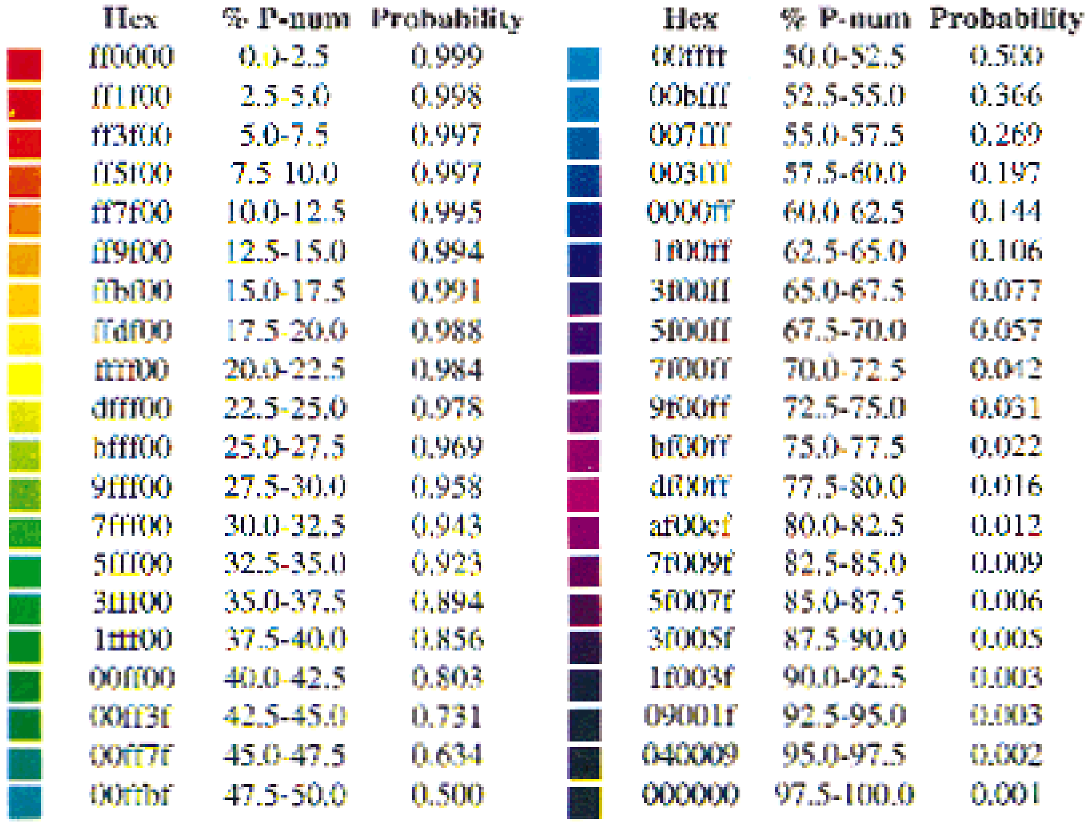
In addition to color, we also use another form of annotation to show how closely two structural models are related to one another. The short line segments that denote base pairs in a structure plot can be thickened to denote base pairs that are conserved in a reference folding. This feature can be used to visualize conformational differences between wild-type and mutant genomes, between predicted alternative foldings, and between the phylogenetic and predicted models of RNA molecules. The application of these approaches to the analysis of several molecules of RNase P RNA and 16S rRNA are given below.

APPLICATIONS

We chose a range of colors that vary fairly smoothly from red through orange, yellow, green, cyan, blue, magenta, and finally black. These colors are used to represent bases or base pairs that are "well-determined" to "poorly determined," respectively. The exact correspondence between color and various "well-determinedness" indices is given in Figure 1. These colors were chosen for their visual appeal. The red, very "well-determined" regions catch the eye. The range of colors gives a good visual difference between pairs of discriminant measures.

Initially, we colored bases in structure plots. In this way, both the base identity and its level of "well-determinedness" could be shown simultaneously. Unfortunately, we found that the colors appeared too faintly in the annotated plots. In addition, the bases are too small to be seen in plots of large foldings. For this reason, we chose to plot colored disks, or dots as we call them, that are roughly the size of the base characters they replace. The resulting plots are visually appealing and informative, even for very large molecules. Base coloring has nevertheless been retained as an option in the annotation programs.

When annotation is based on P-num or S-num, each base is colored according to its P-num or S-num value. We call this *base-dependent* annotation. The P-num or S-num values are scaled linearly from 0 to 1 by dividing by the maximum. The colors are chosen according to a linear scale (Fig. 1). In base-dependent annotation methods, paired bases are not necessarily the same color. In a particular folding, one partner of a base pair might



	Hex	% P-num	Probability		Hex	% P-num	Probability
🟥	ff0000	0.0-2.5	0.999	🟦	00ffff	50.0-52.5	0.500
🟥	ff1f00	2.5-5.0	0.998	🟦	00bfff	52.5-55.0	0.366
🟥	ff3f00	5.0-7.5	0.997	🟦	007fff	55.0-57.5	0.269
🟥	ff5f00	7.5 10.0	0.997	🟦	003fff	57.5-60.0	0.197
🟧	ff7f00	10.0-12.5	0.995	🟦	0000ff	60.0 62.5	0.144
🟧	ff9f00	12.5-15.0	0.994	🟪	1f00ff	62.5-65.0	0.106
🟨	ffbf00	15.0-17.5	0.991	🟪	3f00ff	65.0-67.5	0.077
🟨	ffdf00	17.5-20.0	0.988	🟪	5f00ff	67.5-70.0	0.057
🟨	ffff00	20.0-22.5	0.984	🟪	7f00ff	70.0-72.5	0.042
🟨	dfff00	22.5-25.0	0.978	🟪	9f00ff	72.5-75.0	0.031
🟩	bfff00	25.0-27.5	0.969	🟪	bf00ff	75.0-77.5	0.022
🟩	9fff00	27.5-30.0	0.958	🟪	df00ff	77.5-80.0	0.016
🟩	7fff00	30.0-32.5	0.943	🟪	af00cf	80.0-82.5	0.012
🟩	5fff00	32.5-35.0	0.923	🟪	7f009f	82.5-85.0	0.009
🟩	3fff00	35.0-37.5	0.894	🟪	5f007f	85.0-87.5	0.006
🟩	1fff00	37.5-40.0	0.856	⬛	3f005f	87.5 90.0	0.005
🟩	00ff00	40.0-42.5	0.803	⬛	1f003f	90.0-92.5	0.003
🟩	00ff3f	42.5-45.0	0.731	⬛	09001f	92.5-95.0	0.003
🟦	00ff7f	45.0-47.5	0.634	⬛	040009	95.0-97.5	0.002
🟦	00ffbf	47.5-50.0	0.500	⬛	000000	97.5-100.0	0.001

FIGURE 1. Color annotation used to indicate the propensity of individual nucleotides to participate in base pairs and whether or not a predicted base pair is well-determined. Forty colors that range from red (unusually well-determined) to black (poorly determined) are used. Their hexadecimal values are shown in column 1. The corresponding %P-num is shown in column 2. The %P-num value that is used to annotate each individual nucleotide of the structure plot is calculated from the P-num table generated by the *mfold* package. Its absolute value depends on the energy range that the user selects in creation of the energy dot plot. Probability values shown in column 3 are used to annotate structure with probability information from the Vienna RNA folding package. A double logarithmic scale is used, as described in the text.

pair with only a few other bases in all close to optimal foldings, whereas the other partner might pair with many other bases. There is no a priori reason to expect symmetry.

Structure plots of two molecules of RNase P RNA (Reed et al., 1982; LaGrandeur et al., 1993; Brown, 1998) that were annotated with P-num values can be seen in Figure 2A and B. The annotated plot for *Escherichia coli* (Fig. 2A) shows that this structure is relatively poorly determined. Most dots are colored from light green to blue and some are dark purple. A few dots at the apex of several hairpins are bright red, indicating that nucleotides within these small local regions are well-determined. In contrast, large helices in the annotated plot of the *Sulfolobus acidocaldarius* RNase P RNA (Fig. 2B) are very well-determined and are colored bright red. Additional features throughout the structure are colored orange and yellow, indicating that they, too, are relatively well-determined.

In Figure 2C and D, we show the energy dot plots that correspond to the color annotations given in Figure 2A and B. The energy dot plots give the superposition of all base pairs within 5% of the optimal folding. They contain more information than the annotated structure plots (see Discussion), but are also more difficult to interpret. In each energy dot plot, the optimal folding is represented in black. Suboptimal base pairs are represented in color. Well-determined features can be recognized easily because they are located in clear areas of the plot where few alternative base pairs form. Inspection of the dot plot for *E. coli* RNase P RNA (Fig. 2C) shows that it is poorly determined; a uniform distribution of base pairs is seen at all levels of suboptimality. In contrast, the dot plot for *S. acidocaldarius* RNase P RNA (Fig. 2D) is relatively well-determined. One small domain extending from nt 116 to 193 (located in the center of plot near the diagonal) has virtually no competing base pairs. Similarly, a long helix that pairs the 5′ and 3′ ends of the molecule is also well-determined. These regions correspond to the helices shown in red in Figure 2B. In our own studies, we use both the energy dot plot and the annotated structure plot of each folding prediction to analyze the folding potential of the predicted structure. The energy dot plot gives a good overview of the folding potential of the entire molecule, and the structure plot is used to extract detailed information about specific base paired regions.

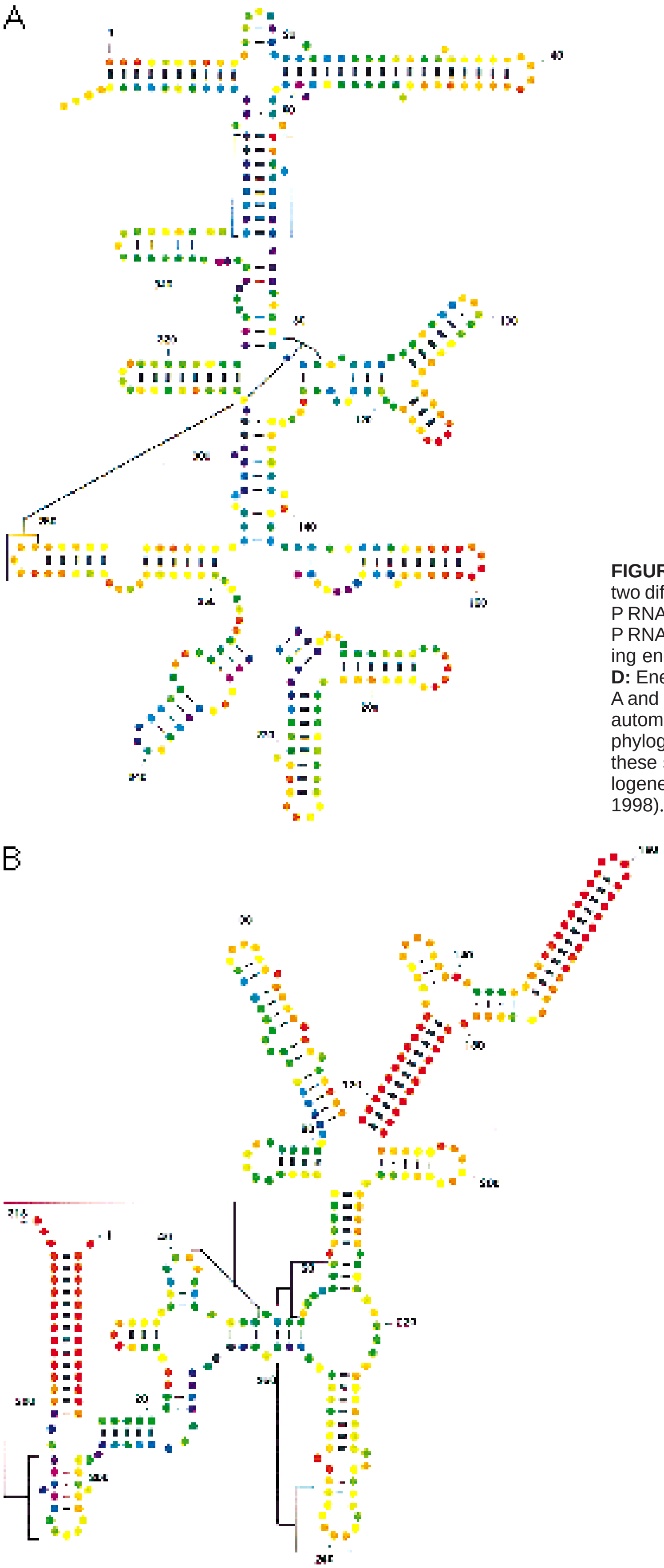


FIGURE 2. Illustration of how well-determined the prediction is for two different molecules of RNase P RNA. **A:** Structure plot for RNase P RNA from *E. coli* annotated with P-num. **B:** Structure plot for RNase P RNA from *S. acidocaldarius* annotated with P-num. **C:** Corresponding energy dot plot for the annotated structure plot shown in part A. **D:** Energy dot plot corresponding to part B. RNA structures shown in A and B are suboptimal foldings. They were selected from a group of automatically generated foldings based on their consistency with the phylogenetic models. Thick lines were used to indicate base pairs in these structures that correspond to base pairs in the published phylogenetic models (Reed et al., 1982; LaGrandeur et al., 1993; Brown, 1998). (*Figure continues on facing page.*)

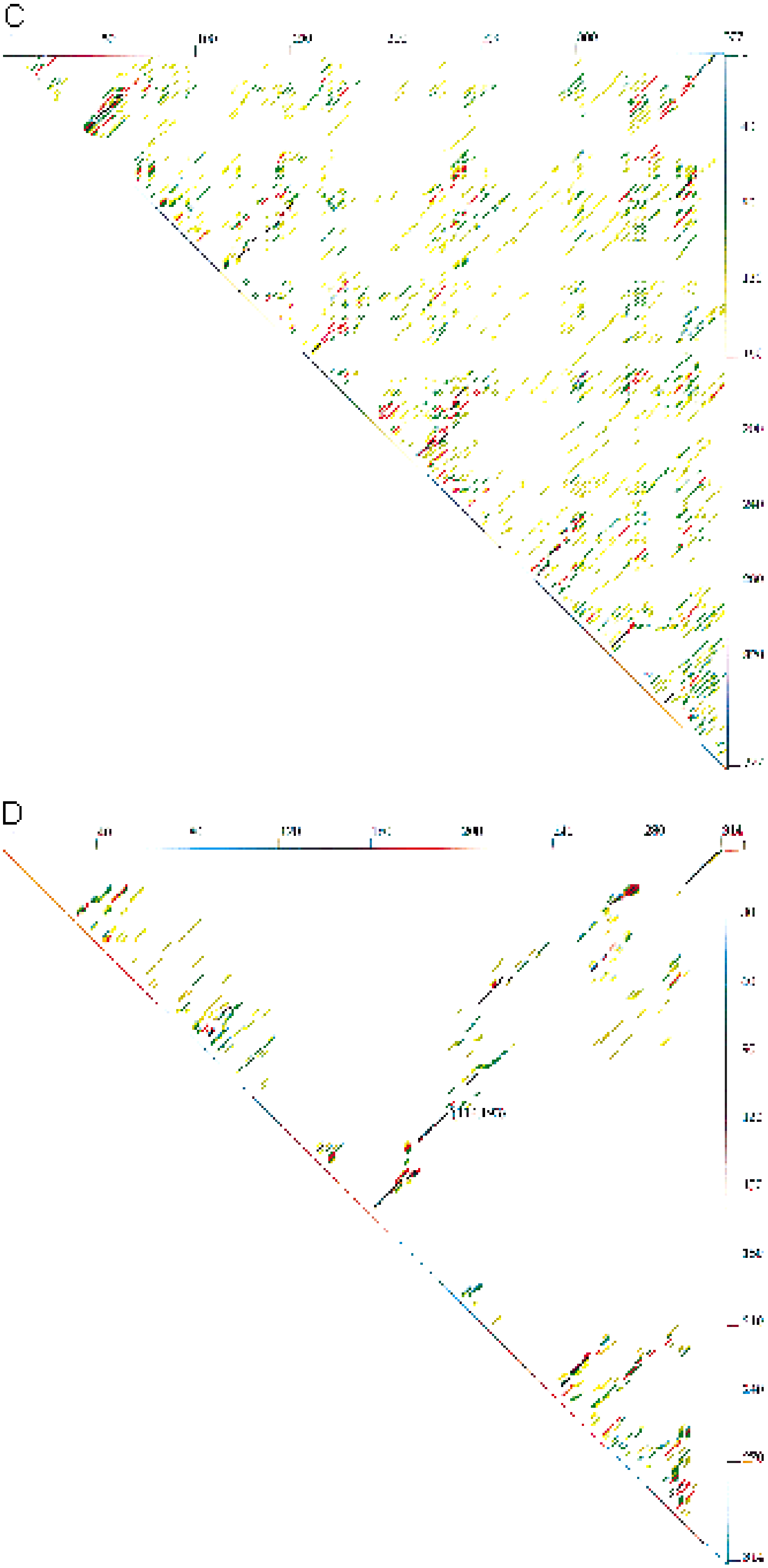


FIGURE 2. (*Continued.*)

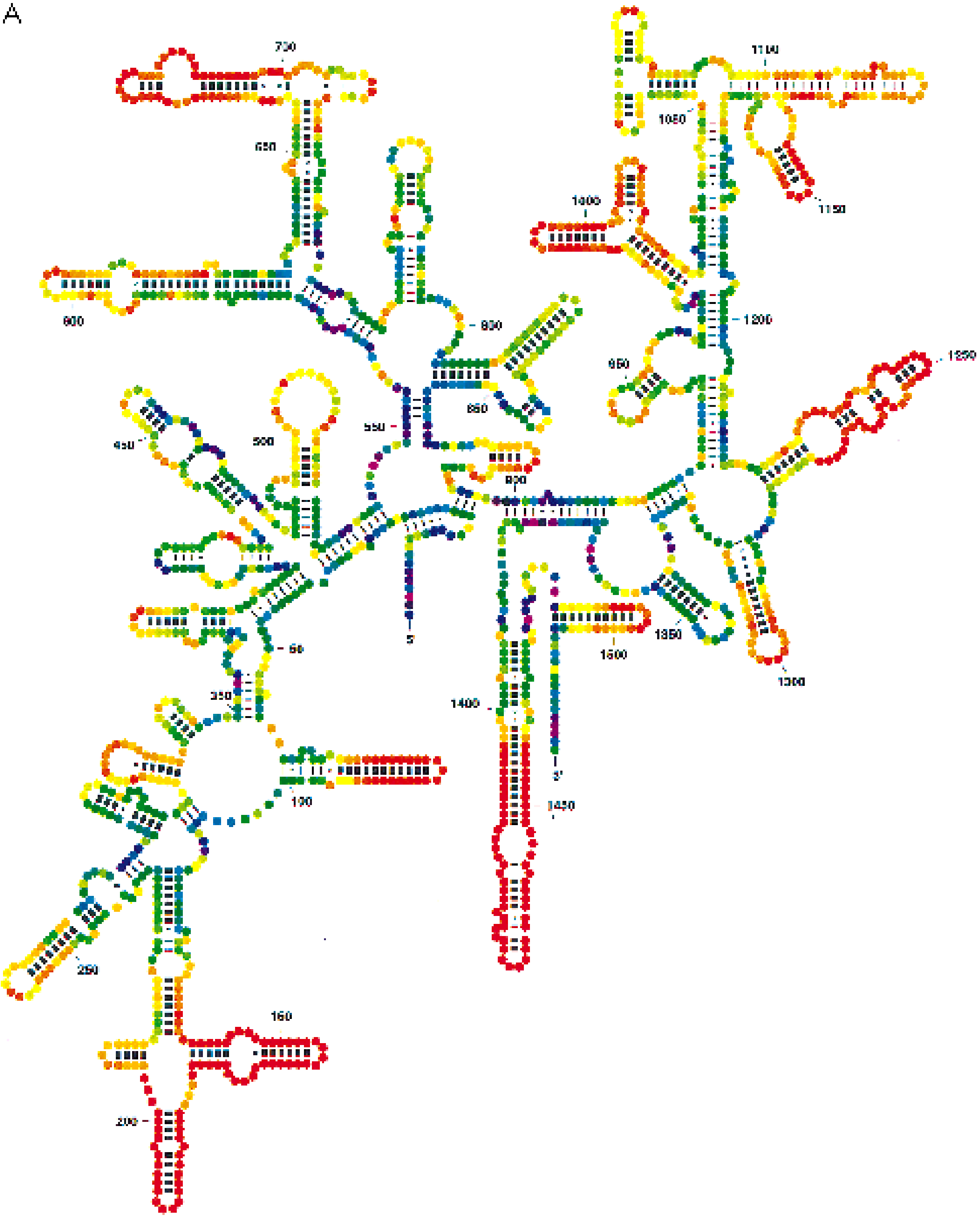


FIGURE 3. (*Figure continues on facing page.*)

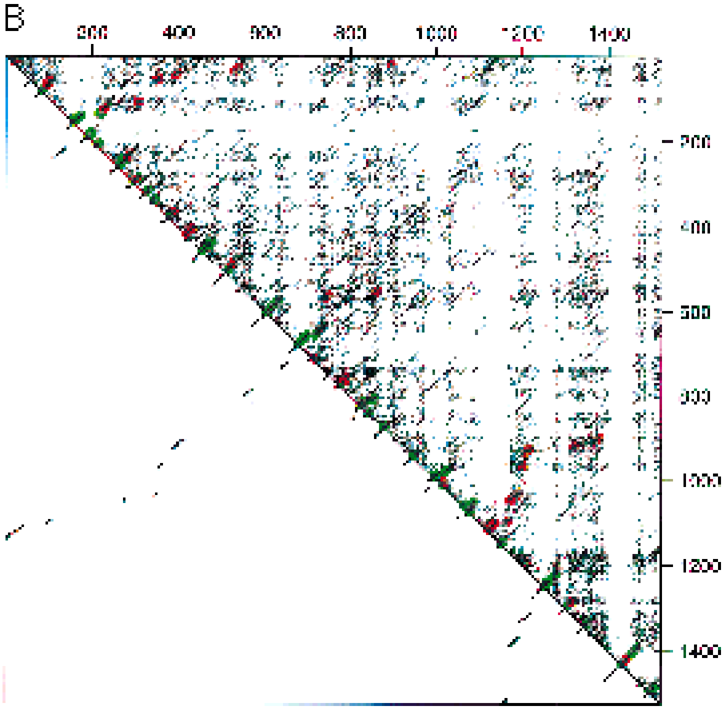


FIGURE 3. Annotation of the phylogenetic model of 16S rRNA from *T. thermophilus*. **A:** Color annotation based on P-num. Annotation is based on a 12-kcal energy dot plot created with version 2.3 of *mfold*. Thick lines are used to show base pairs in the predicted optimal folding of the RNA (not shown) that are also present in the phylogenetic model (Murzina et al., 1988; Gutell, 1994). **B:** 12-kcal overlay energy dot plot. The optimal predicted folding is shown in the lower left triangle. All base pairs within 12 kcal of the optimal folding are shown in the upper right triangle. Red and green lines indicate the position of helices in the phylogenetic model.

Structure comparison annotation is achieved through a program called *ss-compare*. With it, base pairs in a reference structure are represented by thick lines that stand out from the nonconserved base pairs. The annotation can be used to illustrate differences between pairs of alternative predicted foldings, between wild-type and mutant genomes (Jacobson et al., 1998), and between phylogenetically determined versus computer-predicted models of an RNA secondary structure. An example of the latter is shown in Figure 2A and B. In addition to color annotation, the structures have also been annotated with ss-compare to show whether any of the predicted helices are present in the published phylogenetic models for the two RNAs. In the *E. coli* plot, 10 of 16 of the helices that are shown in the plot are also found in the phylogenetic model for this RNA. In the *S. acidocaldarius* plot, 8 of a total of 12 helices are in the phylogenetic model. The structures that are shown are not optimal. They are selected from a group of automatically generated foldings. The selection criterion is the greatest degree of consistency with the published phylogenetic models for these RNAs. Approximately 60% of the predicted helices in the optimal foldings of the two RNAs match the phylogenetic models.

The program ss-compare can also be used to annotate a phylogenetic model of an RNA in order to examine how well individual structures are predicted. This feature is illustrated in Figure 3A, where the phylogenetic model 16S ribosomal RNA from *Thermus thermophilus* (Murzina et al., 1988; Gutell, 1994) has been annotated with color to show how well-determined individual base pairs are, and with thick lines to indicate which structural features are predicted correctly by *mfold*. The corresponding energy dot plot for this RNA is shown in Figure 3B. It is presented as an overlaid dot plot (Zuker & Jacobson, 1995) where predicted helices in the optimal folding are underscored in green if they are also found in the phylogenetic model. Large red lines indicate the position of helices in the phylogenetic model that are absent from the predicted optimal folding. It can be seen that suboptimal helices are located at many of these sites. Both the annotated structure plots (Fig. 3A) and the energy dot plot (Fig. 3B) of 16S rRNA from *T. thermophilus* show that the majority of local hairpins along the outer edges of the structure are extremely well-determined and that all of these structures are predicted correctly. Poorly determined regions are located predominantly within the interior of the molecule, and only some of these helices were predicted correctly by *mfold*.

Annotation can also be *base pair-dependent*. This is especially relevant when annotation is based on box-plot probabilities from the Vienna package. Both bases in a base pair are given a color that corresponds to the probability of that base pair. Single-stranded bases are annotated according to the probability that they are single-stranded. P-num values can also be used in a base pair-dependent way by using an average value, $[P\text{-}num(i) + P\text{-}num(j)]/2$, to annotate both bases, i and j, in a pair. When probabilities are used for annotation, a double logarithmic scale is used to assign colors. For probabilities, p, ranging from 0.999 to 0.5, log(1.0 − *p*) is mapped linearly to the top 20 colors. For probabilities, *p*, in the range from 0.5 to 0.001, log*p* is mapped linearly to the bottom 20 colors. This mapping makes it possible to distinguish easily among both high and low probabilities.

A base pair-dependent annotation of the phylogenetic folding of 16S rRNA for *T. thermophilus* can be seen in Figure 4. Base pair probabilities were calculated with the Vienna RNA folding package using version 1.2.1. The corresponding color annotation is given in Figure 1. Overall, the plots shown in Figure 3A and in Figure 4 bear a striking resemblance to one another. Well-determined hairpins at the edges of the structure are similar in position in both plots. A striking difference is observed, however, within the interior of the two plots. Many long-range helices that are only predicted to be poorly determined by *mfold* are predicted to be completely improbable (shown in black) with the Vienna package. The significance of these observed differences in the prediction by the two algorithms remains to be explored. It is useful to bear in mind that although both software packages are using almost identical energy functions, *mfold* computes what base pairs are *possible* in close to optimal foldings, whereas the Vienna package computes base pair *probabilities*.

The final annotation method used is called *helix dependent*. All the bases in a helix are colored according to the H-num value of the helix. As with P-num and S-num, the numbers are divided by the largest value, and thus range between 0 and 1. No example is shown for this annotation.

DISCUSSION

We have described several new computer programs that have been designed to enable the user to determine the reliability of predicted RNA secondary structures relatively easily. Colored dots are used to provide a simple overview of how well-determined an entire predicted structure and/or a predicted structural domain within a larger folding is. Although the same information can be obtained from dot plot representations of optimal and suboptimal foldings that is created both by *mfold* and by the Vienna package, the dot plot representation of RNA secondary structures is relatively difficult to understand. In addition, it is sometimes difficult to integrate information in the dot plots accurately by eye. Thus, the use of color is helpful even to an experienced user.

Assessing the reliability of a prediction

Published analyses with ribosomal RNAs (Zuker & Jacobson, 1995) and with coliphage Qβ (Jacobson & Zuker, 1993; Jacobson et al., 1998) have shown that well-determined structural features are more likely to be predicted correctly by *mfold*. Another study has examined the reliability of structure prediction using the Vienna program that calculates base pair probabilities for a large number of 16S and 23S rRNAs (Huynen et al., 1997). These authors have similarly shown that "well-determined" predicted structures are more likely to be predicted correctly. That is, "well-determined" structures contain a greater percentage of correct base pairs than do "poorly determined" structures. Their notion of "well-determined" is equivalent to a low *entropy measure*, which is derived from the boxplot base pair probabilities. These studies by other groups have analyzed entire structures. They have not, however, examined portions of secondary structures to see whether some features are more likely to be correct than others. It also remains unknown whether or not base pairs with high probabilities are more likely to be correct, that is, in comparative models.

The fact that some predicted structures or parts of these structures are "well-determined" and others are "poorly determined" may mean nothing more than that the former type of predictions are more reliable. We believe that this phenomenon means more. The analyses with ribosomal RNAs (Zuker & Jacobson, 1995; Huynen et al., 1997) and our own structural analyses of wild-type and mutant coliphage Qβ RNAs (Jacobson et al., 1998) suggest that the relative frequency of predicted alternative conformations of the RNA can reflect physical properties of the RNA. Well-determined 16S rRNA predictions are found primarily among the Archaea, organisms that grow in harsh environments and at high temperature. The structure of rRNAs in organisms that grow in these environment are likely to be optimized both to fold efficiently and to be unusually stable. In coliphage RNAs, experimental studies show that well-determined structural domains correspond to domains within the RNA that are unusually stable.

In addition to well-determined structures, the prediction of poorly determined structures may provide insight into regions of potential structural plasticity within an RNA molecule. In coliphage RNAs, two cases have been found where competing alternative conformers are found in regions of the RNA that are predicted to be poorly determined by computer modeling (Jacobson & Zuker, 1993; Jacobson et al., 1998). The analysis of these RNAs suggests that stable structural domains lie interspersed among regions where greater structural plasticity is observed. However, the observed correspondence between poorly determined structural domains and real structural plasticity for RNA coliphage RNA may not be true of other RNAs. For example, the entire predicted structure of many ribosomal RNAs is poorly determined. Although there is growing recognition that both protein chaperones and small RNAs may contribute to the proper folding of these RNAs within the living cell (Konings & Gutell, 1995), the analysis of the conformation of 16S rRNA from *E. coli* with chemical and enzymatic probes has shown that the structure

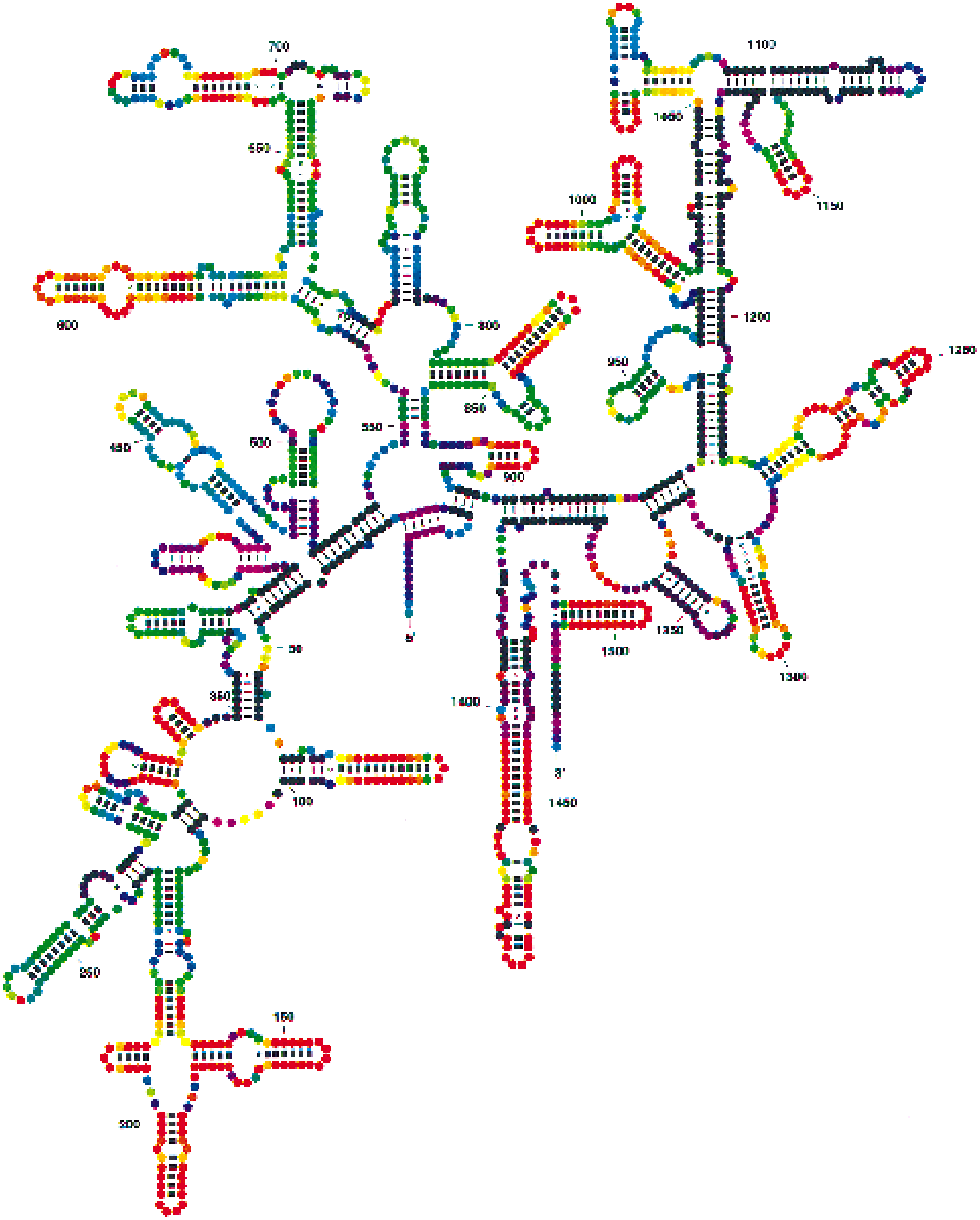


FIGURE 4. Color annotation of the phylogenetic model for 16S rRNA of *T. thermophilus* based on probability. The probability for each base pair was calculated using version 1.2.1 of the Vienna RNA folding package. This program version uses energy rules that are identical to those used in version 2.3 of *mfold*. The predicted folding is almost indistinguishable from the predicted optimal folding that is generated with version 2.3 of *mfold*.

of this RNA is unique and consistent with the phylogenetic model for this RNA (Noller, 1984; Murzina et al., 1988; Gutell, 1994). Because the predicted structure for this entire RNA is poorly determined (Zuker & Jacobson, 1995), it is clear that the physical interpretation of poorly determined structural features that are obtained by computer modeling may be complex and may always require experimental verification. In the case of the *E. coli* rRNA, the poorly determined prediction might indicate that the molecules are easily misfolded in solution.

Merits of annotation of folding versus direct dot plot analysis

Although color annotation of a given folding provides an easy overview of many features and properties of a predicted folding, it does not substitute entirely for the direct examination of the dot plot output. In many instances, more than one optimal folding is found for the same RNA sequence. Such structures are identified most readily in dot plots. It is also useful to examine close competing alternative structures in poorly determined regions of the plot. This is particularly useful when independent structural information is available that can be used to select likely structural candidates. Common structures in related RNA sequences are often found among suboptimal structures. Other types of auxiliary structural information can also often be used to identify likely suboptimal foldings in regions that are only moderately well-determined.

Visual inspection of dot plots can also lead to the prediction of separate folding domains that do not interact with one another, as was the case in our analysis of the wild-type Qβ genome (Jacobson & Zuker, 1993). Color annotation on predicted structures does not tell us whether the alternative base pairs formed by bases in a particular domain are *intradomain* or *extradomain*. Only the dot plot can give us this information.

Future directions

The "Vienna group" has developed several additional terms, based on statistical mechanics and information theory, that can be used to describe how well-determined a predicted structure or subsection thereof is. These include a "well-definedness" measure, $d(i)$ and an entropy measure, $S(i)$, for every base i in the RNA (Huynen et al., 1997). These descriptors can also be used to annotate predicted structures with color in the manner that we have described. These descriptors are neither better nor worse than the base pair probabilities, but they convey different information. It still remains to be seen whether these alternative approaches will provide additional insight into RNA structure that would be useful to the experimentalist.

Although base annotation based on P-num has already proved extremely useful in several analyses of RNA structure (Jacobson et al., 1998; Palmenberg & Sgro, 1998), the approach relies heavily on the basic reliability of the RNA structure prediction itself and is therefore very much dependent on the energy rules that are used in RNA structure prediction. Although the prediction of local hairpins now appears to be quite reliable, multibranch junctions and long-range interactions remain problematic. Work is currently underway to implement the helix stacking feature in multibranch junctions that is described in Walter et al. (1994). It is expected that the implementation of this feature will improve the reliability of long-range prediction. In addition, it may alter the "well-definedness" of individual structural features.

METHODS

Two programs have been created to achieve the desired color annotation. The first one is called *bp-annotate*. This program annotates the PostScript structure plots that are created by the *mfold* package, or the PostScript secondary structure plots produced by the XRNA (Weiser & Noller, 1995) program. Either the bases are colored or the bases are replaced by colored dots. Both linear and double logarithmic scaling can be used. Two input files are needed; the unannotated PostScript file and an annotation file that we call the *ann* file because it ends with the suffix ".ann." This annotation file can contain either P-num, H-num, or probability information. A number of small auxiliary programs were created to generate this from different types of input information. The one most worth mentioning is called the *boxct2ann* program. It combines information from a probability boxplot produced, for example, by the Vienna package, and the *ct* file for the secondary structure to be annotated. The boxplot contains numbers, p_{ij}, for every possible base pair. These are the probabilities of designated base pairs. The *ct* file, or connection table file, is a common format for describing the bases and base pairs in an RNA secondary structure. The resulting *ann* file is a vector of probabilities, p_i, where i ranges over all bases of the sequence. For base pairs $i{\cdot}j$, $p_i = p_j = p_{ij}$. If i is single-stranded, then p_i is the probability that the base is single-stranded. In this case, $p_i = 1 - \sum_{j \neq 1} p_{ij}$.

The second program, *ss-annotate*, annotates the *ss* files that are used as input to the XRNA program. The figures can then be manipulated interactively by running the XRNA program. This is useful both in creating high-quality plots, and for extracting detailed information about individual structural features from the annotated plots. The latter feature is particularly helpful in the analysis of predicted structure of very large sequences. The same color schemes and linear versus double logarithmic scaling are used in both annotation programs.

Structure comparison annotation is achieved through a program called *ss-compare*. It requires the *ss* files that are used as input to the program XRNA. In addition to an *ss* file, the program also requires the *ct* file for another reference structure on the same RNA. The base pairs in the *ss* file that are conserved in the reference structure are represented by thick lines that stand out from the nonconserved base pairs.

The programs *bp-annotate* and *boxct2ann* were written in Fortran. The programs *ss-compare* and *ss-annotate* were written in the C programming language. All of the software can be obtained from M. Zuker and is available at ftp://snark.wustl.edu/pub. Energy dot plots and annotated structure plots are also available for all structure predictions created with the *mfold* web server at this site: (http://www.ibc.wustl.edu/~zuker/rna/form1.cgi).

ACKNOWLEDGMENTS

We thank Bryn Weiser and Harry Noller for making the program XRNA available to us. We also thank Bryn Weiser for his support during the implementation of the new color annotation in XRNA. Jason Ginchereau created the *ss* file annotation and comparison programs. Darrin Stewart created the current program for displaying energy dot plots. The research was supported by grant MCB316501 from the National Science Foundation and National Institutes of Health (NIH) grant 48702 to A.B.J. M.Z. has been supported by grant GM54250 from the NIH.

Received January 21, 1998; returned for revision February 27, 1998; revised manuscript received March 24, 1998

REFERENCES

Brown JW. 1998. The ribonuclease P database. *Nucleic Acids Res 26*:351–352.

Gutell RR. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res 22*:3502–3507.

Hofacker IL, Fontana W, Stadler PF, Bonhöffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie 125*(2):167–188.

Huynen M, Gutell RR, Konings DA. 1997. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol 267*:1104–1112.

Jacobson AB, Arora R, Priano C, Lin CH, Zuker M, Mills D. 1998. Structural plasticity in RNA and its role in the regulation of protein translation in coliphage Qβ. *J Mol Biol 274*:589–600.

Jacobson AB, Zuker M. 1993. Structural analysis by energy dot plot of a large mRNA. *J Mol Biol 233*:261–269.

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA 86*:7706–7710.

Jaeger JA, Turner DH, Zuker M. 1990. Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol 183*:281–306.

Konings DA, Gutell RR. 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA 1*:559–574.

LaGrandeur TE, Darr SC, Haas ES, Pace NR. 1993. Characterization of the RNase P of *Sulfolobus acidocaldarius*. *J Bacteriol 175*:5043–5048.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers 29*:1105–1119.

Murzina NV, Vorozheykina DP, Matvienko NI. 1988. Nucleotide sequence of *Thermus thermophilus* HB8 gene coding 16S rRNA. *Nucleic Acids Res 16*:8172.

Noller HF. 1984. The structure of ribosomal RNA. *Annu Rev Biochem 53*:119–162.

Palmenberg AC, Sgro JY. 1998. Topological organization of picornaviral genomes: Statistical prediction of RNA structural signals. *Seminars in Virology 8*:231–241.

Reed RE, Baer MF, Guerrier-Takada C, Donis-Keller H, Altman S. 1982. Nucleotide sequence of the gene encoding the RNA subunit (M1 RNA) of ribonuclease P from *Escherichia coli*. *Cell 30*:627–636.

Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH, Zuker M. 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci USA 91*:9218–9222.

Weiser B, Noller HF. 1995. *XRNA: Auto-interactive program for modeling RNA*. The Center for Molecular Biology of RNA, Santa Cruz, California: University of California. Internet: ftp://fangio.ucsc.edu/pub/XRNA.

Williams AL, Tinoco I Jr. 1986. A dynamic programming algorithm for finding alternate RNA secondary structures. *Nucleic Acids Res 14*:299–315.

Zuker M. 1989a. On finding all suboptimal foldings of an RNA molecule. *Science 244*:48–52.

Zuker M. 1989b. The use of dynamic programming algorithms in RNA secondary structure prediction. In: Waterman MS, ed. *Mathematical methods for DNA sequences*. Boca Raton, Florida: CRC Press, Inc. pp 159–184.

Zuker M. 1994. Prediction of RNA secondary structure by energy minimization. In: Griffin AM, Griffin HG, eds. *Computer analysis of sequence data, part II, vol 25*. Totowa, New Jersey: Humana Press, Inc. pp 267–294.

Zuker M, Jacobson AB. 1995. "Well-determined" regions in RNA secondary structure prediction. Analysis of small subunit ribosomal RNA. *Nucleic Acids Res 23*:2791–2798.

RNA

A PUBLICATION OF THE RNA SOCIETY

Using reliability information to annotate RNA secondary structures.

M Zuker and A B Jacobson

RNA 1998 4: 669-679