Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 9408–9412, October 1993
Biochemistry

Reproducing the three-dimensional structure of a tRNA molecule from structural constraints

FRANÇOIS MAJOR*, DANIEL GAUTHERET†, AND ROBERT CEDERGREN†

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; and †Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada

Communicated by Elvin A. Kabat, July 21, 1993

ABSTRACT **The three-dimensional structure of yeast tRNA^{Phe} was reproduced at atomic resolution with the automated RNA modeling program MC-SYM, which is based on a constraint-satisfaction algorithm. Structural constraints used in the modeling were derived from the secondary structure, four tertiary base pairs, and other information available prior to the determination of the x-ray crystal structure of the tRNA. The program generated 26 solutions (models), all of which had the familiar "L" form of tRNA and root-mean-square deviations from the crystal structure in the range of 3.1–3.8 Å. The interaction between uridine-8 and adenosine-14 was crucial in the modeling procedure, since only this among the tertiary pairs is necessary and sufficient to reproduce the L form of tRNA. Other tertiary interactions were critical in reducing the number of solutions proposed by the program.**

Since the spatial disposition of a molecule is generally thought to be responsible for its activity, the search for methods having the ability to determine or model three-dimensional structures has been intense. X-ray crystallography, the most respected method of structure determination, has provided a three-dimensional structure of only one biologically active RNA, that of tRNA (1, 2). The increasing number of important small RNAs underlines the need for new structure determination methods. Recently we reported the application of an algorithm based on constraint satisfaction to the problem of macromolecular modeling (3). This algorithm exhaustively searches conformational space such that all models consistent with a given set of input constraints are produced; thus, contrary to such optimization techniques as energy minimization, the algorithm is not subject to the problem of local minima.

In the MC-SYM program, three-dimensional structures of RNA are produced by the stepwise addition of nucleotides having one or several different conformations to a growing oligonucleotide model. This method leads to the formation of a tree structure where each intermediate solution is a node and the tree size or the theoretical number of terminal nodes is the product of the number of nucleotide conformations allowed at each position. To reduce the combinatorial explosion of this formulation, a limited number of precalculated conformations are used to sample nucleotide conformational space and intermediate solutions which do not satisfy all constraints are pruned from the tree. Also, the "lazy" evaluation feature of the algorithm allows for conformational searches using only the positions of atoms actually defined in the constraints. Once a solution is found the remaining atoms are grafted onto the structure depending on the precalculated conformation of the given nucleotide.

Experimental or theoretical information on RNA structure can be entered in the MC-SYM search protocol either by defining structural constraints required in the final solution or by judiciously selecting possible nucleotide conformations. This modeling procedure has been applied to a series of small loop structures, but to be useful much larger structures must be confronted. The work reported here uses an implementation of the program in C++ language which runs 2 orders of magnitude faster than the original Miranda code and thereby permits modeling of longer RNA molecules.

The precalculated sets of nucleotide conformations were derived from a nucleic acid structural database assembled from structures determined by x-ray crystallography and NMR spectrometry (4). Extensive testing of different sampling techniques demonstrated that sufficient modeling precision could be obtained by using 30 and occasionally 50 different conformations to represent unconstrained nucleotides in a structure (the *All30* and *All50* conformational sets). Structural information on a nucleotide, such as its sugar pucker or glycosidic torsion angle or whether it is base paired or stacked, is incorporated into the modeling protocol by conformational sets derived from sampling nucleotides in the database having these characteristics. The conformational sets used in the modeling of the tRNA^{Phe} were A-helical nucleotides (*TypeA*, one conformation), B-helical nucleotides (*TypeB*, one conformation), stacked nucleotides (*StkAA*, three conformations), and *All30* or *All50*.

Modeling with MC-SYM involves the writing and testing of different input scripts describing the available structural knowledge. Scripts that were written to evaluate the effect of the order of constraints on execution time have shown that it is advantageous to use the most limiting constraint early in the protocol, to avoid the generation of large numbers of branches near the root of the search tree. However, more important for the present study was the evaluation of how the quantity and the value or precision of constraints in a script would affect the number and quality of solutions. Here the best strategy has been to identify the most constraining or limiting script that yields solutions. This approach has the advantage that fewer solutions are produced and subjected to evaluation. The quality of solutions was judged by the root-mean-square (rms) deviation from the known crystal structure of tRNA^{Phe}.

We have chosen to model the tRNA molecule because it has been the benchmark for RNA modeling and structural studies for some time. Constraints for the prediction of the tRNA structure have been formulated from structural inferences available prior to 1970. They include the cloverleaf base-pairing pattern, some base and helical-region stacking, tertiary base interactions, most of which were predicted by Levitt (5), and a series of steric constraints. The tertiary interactions can also be detected by covariation analysis of aligned tRNA sequences (ref. 6; M. Turcotte and R.C., unpublished data). Another important constraint derives from the fact that the combination of loop nucleotide conformations must permit loop closure. A full list of the constraints and the conformational sets for each nucleotide are given in Table 1 and Fig. 1, respectively.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Constraints used in modeling

nt 1	nt 2	Constraint	Upper bound, Å
9	10	O3′–P	4.5
16	22	Feasible	3.5
17	22	Feasible	3.5
18	22	Feasible	3.5
19	22	Feasible	3.5
19	56	N1–N3	4.0
20	22	Feasible	3.5
21	22	O3′–P	3.5
32	34	Feasible	2
33	34	O3′–P	3.25
34	31	Feasible	2
35	31	Feasible	2
36	31	Feasible	2
37	31	Feasible	2
38	31	Feasible	2
46	49	Feasible	2
47	49	Feasible	2
48	49	O3′–P	4.5
48	15	N3–N1	4
56	58	Feasible	2
57	58	O3′–P	2
59	61	Feasible	2
60	61	O3′–P	2
All		C1′–C1′	3.5*
All		P–P	3.5*
All		PSE–PSE	2.5*
All		P–C1′	2.5*
All		P–PSE	2.5*
All		C1′–PSE	2.5*
All		O2′–O4′	2.0*
All		O3′–C5′	2.0*
All		P–C3′	2.0*

Numbers in the first two columns represent the nucleotide position in tRNA; "all" refers to spatial constraints used for all nucleotides or between nucleotide pairs. Numbers and letters in the constraint column refer to the atoms involved in the constraint. In addition, O3′–P stands for backbone closure constraints, and PSE, for a pseudoatom which was used to approximate nitrogen bases. The pseudoatom is at the geometric center of the six-membered ring in the case of either a purine or a pyrimidine. The "feasible" constraint is a look-ahead feature which determines whether the number of nucleotides remaining to be modeled in a loop is sufficient to close the loop within the upper bound. All information here can be transformed automatically into an MC-SYM script.
*The lower bound.

The cloverleaf pattern of base pairing predicts the presence of four helical stems: the acceptor, the anticodon, the D, and the T stems; the gross structural shape of this molecule is determined by the spatial relation of these stems. Examination of only the helical nucleotides reveals that the disposition of uridine-7 (U7) and G10 (see Fig. 1) is of utmost importance: U7 forms a base pair with A66, which, by virtue of its stacking interaction with G65, establishes the relative orientation of the acceptor stem and the T stem. G10 determines the disposition of the anticodon and D stems, since G10 pairs with C25, which in turn stacks with C26 and C27.

The only tertiary constraint in this region is the U8–A14 interaction, whose presence was inferred from a crosslink between U8 and C13 (7). With A14 as the reference nucleotide, U8 was positioned by using all base-pairing possibilities—Watson–Crick, reverse Watson–Crick, Hoogsteen, and reverse Hoogsteen—thus producing 4 × 30 = 120 (from the *All30* conformational set) dinucleotide structures. Nucleotides G10 to C13 were added with the *TypeA* conformational set and A9 was connected to U8 by using the *All30* set. A loop closure constraint was imposed between the 3′ oxygen of A9 and the phosphorus of G10. However, a closure constraint of up to 5 Å led to no solutions. Since relaxation of the loop closure constraint produced no solutions, the number of conformations allowed for U8 and A9 was augmented by the use of the *All50* sets. Solutions were obtained only in the case where U8 was relaxed. When a 4.5-Å loop closure constraint was allowed, three solutions were generated which had the L-type folding typical of tRNA and 2.3- to 3.4-Å rms deviation from the crystal structure (8). Relaxing the constraints even more by applying *All50* to U8 and A9 simultaneously and further expanding the loop closure constraint to 5.0 Å gave only 12 solutions having rms deviations in the range of 2.3–8.4 Å; the 9 additional solutions were more distant from the crystal structure. Interestingly, even though all possible base interactions for U8 and A14 were tested, the three solutions produced contain a reverse Hoogsteen base pair as in the tRNA crystal structure.

Improvements to our conformational set sampling and steric collision avoidance techniques prompted us to remodel the anticodon loop and T loop (3, 4). A script where five nucleotides were stacked on the 3′ side of the stem and two on the 5′ side with a loop closure constraint of 3.25 Å gave two solutions for the anticodon loop, both of which were within 1.2-Å rms deviation of the crystal structure. Increasing the loop closure to 5 Å produced 65 solutions varying between 1.1 and 2.1-Å rms deviation from the crystal structure. The T loop was modeled with the information in Fig. 1, and all types of pairing between U54 and A58 were tested. The first, highly limiting script yielded no solutions because of a collision between the phosphorus atoms of C60 and C61. Solutions were obtained when the extended conformational set *All50* was applied to either C60 (four solutions) or U59 (three solutions). In the first case, the four solutions had 2.0-, 2.4-, 2.6-, and 3.5-Å rms deviations from the crystal structure, and in the case of the augmented conformational sampling at U59, 1.8–3.5 Å. Solutions below 3 Å contained the U54•A58 reverse Hoogsteen pair present in the crystal structure, whereas the other solutions contained a reverse Watson–Crick pair.

The D and extra loops were modeled together by taking advantage of two inter-loop pairs, G19•C56 and G15•C48 (5). Modeling with any of the two possible base pairs between G15 and C48 led to no solutions even when a loop closure constraint of 10 Å was used between C48 and C49. After visualization of some of the unacceptable solutions and the crystal structure, it became evident that the orientation of G15 did not allow for a solution. Since both tertiary interactions are between nucleotides that are distant in the primary structure, the inability to satisfy the constraint is most likely due to the additive imprecision of the conformational sets as the polynucleotide chain is lengthened. Consequently, we decided to model these interactions by introducing distance constraints rather than precise base-pair geometries. In the present case a distant constraint of 4 Å between nitrogen 1 of G15 and nitrogen 3 of C48 was introduced in the script. The most limiting script yielding solutions involved the use of *TypeA* for G43, A44, and G45; the use of *All30* for G46, U47, and C48; and augmenting the flexibility of G26 and C27 at the junction of the D and anticodon stems with the *StkAA* set. With a loop closure constraint of 4.5 Å, two solutions were produced, which differed only in the arrangement of the T loop. Interestingly, these two T-loop models have a reverse Hoogsteen interaction between U54 and A58, and although the distance constraint introduced between G15 and C48 allowed solutions, none reproduced the reverse Watson–Crick interaction of the crystal structure.

To complete the tRNA structure, the remaining D-loop nucleotides were added in a straightforward manner using a distance constraint of 4 Å between C19 and C56. A loop closure of 3.25 Å between A21 and G22 produced 26 solutions

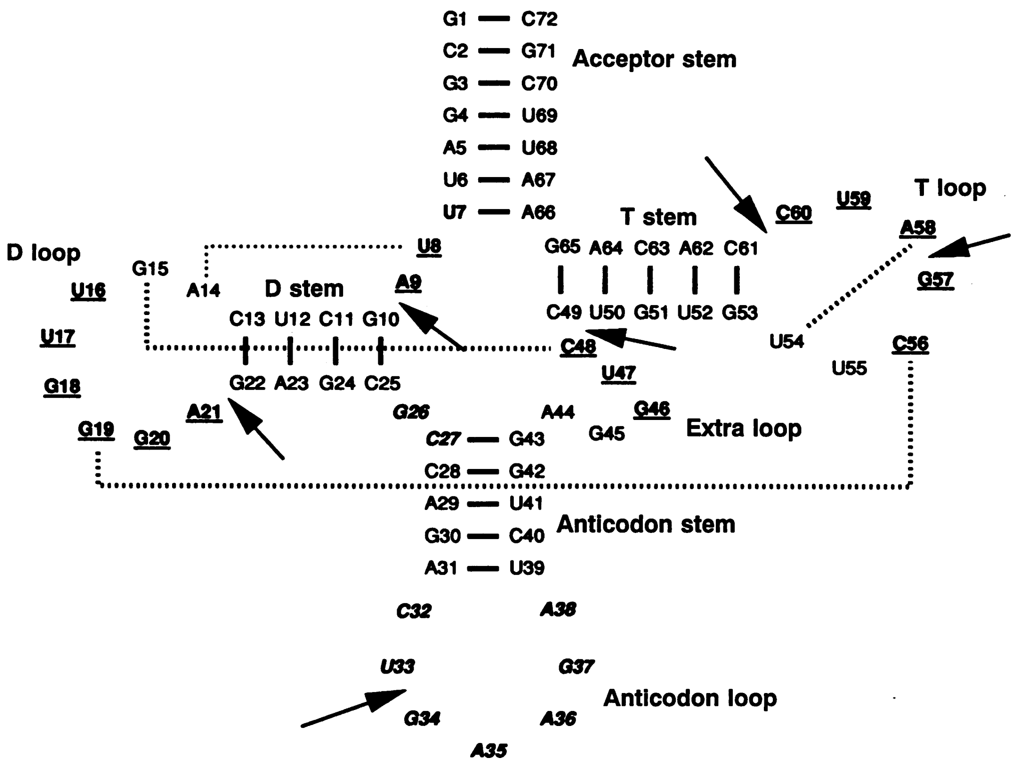

FIG. 1. Structural information used for the prediction of the three-dimensional structure of tRNA. Nucleotides in normal font were assigned the *TypeA* conformational set. Nucleotides in bold were assigned the *TypeB* set. Nucleotides in bold/italics were assigned the *StkAA* conformational set, and those in bold/underlined the *All30*, except for U8 and U59, which were assigned the *All50* set. Bold lines indicate double-helical base pairing. Dotted lines indicate other base pairings. Arrows indicate loop closure constraints. All modeling was performed on the tRNA sequence where modified nucleotides were replaced by their metabolic parent. The number of possible constructions as determined by the product of all conformational set sizes used in the model is 7 × 10^{29}.

containing the two solutions for the anticodon loop and 13 for the D loop. The 13 D-loop models differ from the crystal structure in the range of 4.5- to 7.0-Å rms deviation and are distinctly less precise than the previous domains. The high rms deviations are due to the five consecutive nucleotides U16 to G20 having extended conformations in the crystal structure that are poorly represented in our conformational sets. A summary of the predictions is shown in Table 2, where it can be seen that only one of the original three solutions for the helical core survives the loop modeling process.

The final structures were subjected to 200 iterations of steepest-descent energy minimization, but the G19·C56 and G15·C48 pairs were not reproduced. However, after 3000 iterations of a quasi Newton–Raphson method (14) with fixed atoms in the helices and three distance constraints per base pair, the correct geometry of the G15·C48 base pair was established. As for the known triples G18·U55·A58, C13·G22·G46 and A9·U12·A23, which were not used as constraints in modeling, none was reproduced even though all bases were spatially close to their respective partners. All 26 solutions had the familiar L shape of the tRNA^{Phe} crystal structure. The rms deviations from the crystal structure were in the range of 3.1–3.8 Å. To place this number in context, the rms deviations between the phosphorus atoms in the crystal structures of tRNA^{Phe}, tRNA^{Asp} (8) (entries 3tra and 4tra in the Brookhaven Protein Databank), and our best model were determined: tRNA^{Asp} vs. tRNA^{Phe}, 2.63 Å; tRNA^{Phe} vs. model, 2.90 Å; tRNA^{Asp} vs. model, 2.95 Å. Previous tRNA modeling attempts involving the representation of nucleotides by pseudoatoms and either a distance geometry (9) or a molecular dynamics (10) treatment lack the atomic resolution necessary to obtain detailed comparisons with the crystal structure.

Table 2. Summary of solutions

Fragment type	Fragment	No. of models	rms range, Å
Stems	Acceptor	1	0.69–0.69
	D	1	0.99–0.99
	Anticodon	1	0.75–0.75
	T	1	0.61–0.61
Loops	D	13	4.45–7.02
	Anticodon	2	1.43–1.43
	Variable	1	2.74–2.74
	T	1	2.46–2.46
Hairpin loops	D	13	3.54–5.26
	Anticodon	2	1.17–1.17
	T	1	1.76–1.76
	tRNA core	3	2.32–3.39
	tRNA	26	3.12–3.84

The model having the lowest rms deviation from the crystal structure is shown superimposed on the tRNA^{Phe} structure in Fig. 2. Fig. 2*A* compares the two crystal structures and our model. Among the salient features of this model is the high precision obtained in predicting the angle of the junction between the two coaxial helical complexes. Close examination of the model, however, shows that the loop nucleotides are the most distant from the crystal structure. On the other hand, the stem regions and the anticodon stem–loop are well within the conformational space of the corresponding regions in the crystal structure. The rms deviation of each nucleotide is shown in Fig. 3. The loop regions and particularly the D loop stand out because of their high rms deviations. After modeling by regions the entire molecule was modeled with the complete script on a Sun Sparcstation 630/4 MP; the 26 all-atom structures were produced in 19.9 min of central-processing-unit time.

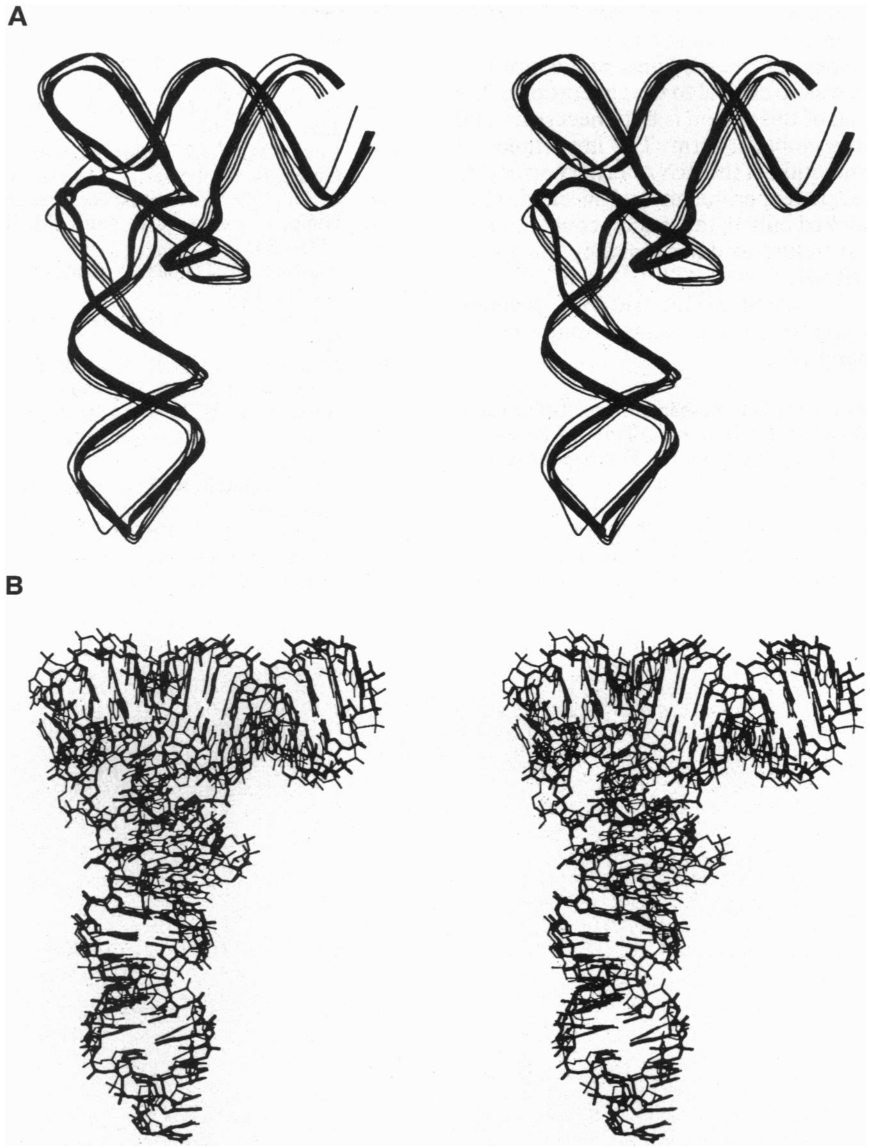

FIG. 2. Stereoviews. (*A*) Overlap of ribbon backbone. Our model is in the wide line; tRNA^{Phe} is represented by a three-line ribbon, and tRNA^{Asp} by a one-line ribbon. (*B*) Superimposition of the all-atom-without-hydrogen representation of the modeled tRNA^{Phe} in light lines and the tRNA^{Phe} in bold lines.

The overall correspondence between the modeled structure and the crystal structure is very encouraging especially when the following factors are considered: (*i*) the crystal structure is subject to crystal packing artifacts (11)—that is, the solution structure could differ from this structure; (*ii*) the x-ray structure is itself a consensus structure; and (*iii*) the structure modeling and minimizations were done in the absence of modified nucleotides which could affect the conformations. In the case of the tRNA^{Phe} structure, the "Y" base is likely to be a major structural determinant in the anticodon loop (12).

Even though the crystal structure was not used explicitly in the modeling, its availability was definitely important, especially in the process of defining the final script. In fact, only a known structure could be of use to us, since our goal was to validate the method and to identify protocols and parameters which produced the highest quality structures, so that this methodology could be generalized to other RNAs. An unknown structure would have been more complicated to model because of the lack of objective criteria for evaluation, although visual inspection and minimum energy levels can be useful in this process. It was reassuring to observe that constraint relaxation after finding solutions with a given

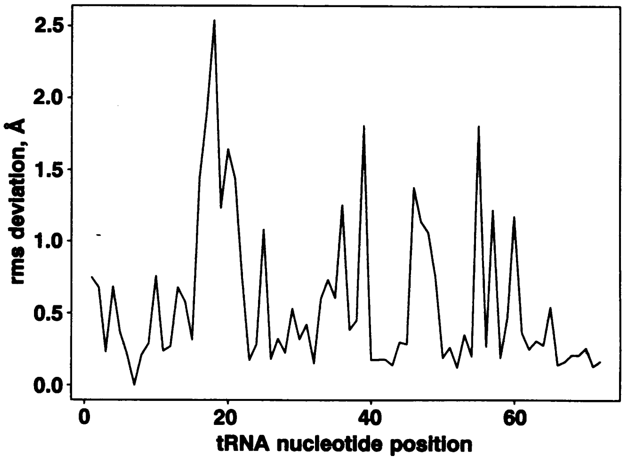

FIG. 3. The rms deviation from the tRNA^{Phe} crystal structure of each position of the modeled tRNA.

script gave solutions which were generally more distant from the crystal structure than the original solutions.

Another surprising aspect of the modeling procedure was how the U7–A14 region was so critical to the generation of the L form. Proper modeling of this region is both necessary and sufficient to produce the global L form. The importance of this region in the determination of the tRNA tertiary structure was also shown by recent experiments of Pan *et al.* (13), where circular tRNA nicked only in this region could not fold to the correct tertiary structure as determined by the ability of Pb^{2+} to cleave the tRNA.

The precision and the speed of the MC-SYM program suggest that RNAs having fewer constraints or more nucleotides could now be handled.

This work was funded by the Medical Research Council of Canada. R.C. is a fellow of the Canadian Institute for Advanced Research. F.M. holds a postdoctoral fellowship from the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche.

1. Quigley, G. J., Seeman, N. C., Wang, A. H. J., Suddath, F. L. & Rich, A. (1975) *Nucleic Acids Res.* **2,** 2329–2341.
2. Moras, D., Comarmond, M. B., Fischer, J., Weiss, R., Thierry, J. C., Ebel, J. P. & Giege, R. (1980) *Nature (London)* **288,** 669–674.
3. Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E. & Cedergren, R. (1991) *Science* **253,** 1255–1260.
4. Gautheret, D., Major, F. & Cedergren, R. (1993) *J. Mol. Biol.* **229,** 1049–1064.
5. Levitt, M. (1969) *Nature (London)* **224,** 759–763.
6. Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. & Stormo, G. D. (1992) *Nucleic Acids Res.* **20,** 5785–5795.
7. Ninio, J., Favre, A. & Yaniv, M. (1969) *Nature (London)* **223,** 1333–1335.
8. Westhof, E., Dumas, P. & Moras, D. (1988) *Acta Crystallogr.* **44,** 112–123.
9. Hubbard, J. M. & Hearst, J. E. (1991) *Biochemistry* **30,** 5458–5465.
10. Malhotra, A., Tan, R. K.-Z. & Harvey, S. C. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 1950–1954.
11. Heinemann, U. (1991) *J. Mol. Struct. Dyn.* **8,** 801–811.
12. Houssier, C. & Grosjean, H. (1985) *J. Mol. Struct. Dyn.* **3,** 387–408.
13. Pan, T., Gutell, R. R. & Uhlenbeck, O. C. (1991) *Science* **254,** 1361–1364.
14. Jacoby, S. L. S., Kowalik, J. S. & Pizzo, J. T. (1972) *Iterative Methods for Nonlinear Optimization Problems* (Prentice–Hall, Englewood Cliffs, NJ).