J. Biochem. 132, 211–215 (2002)

Classification of RNA Structures Based on Hydrogen Bond and Base-Base Stacking Patterns: Application for NMR Structures[1]

Akitsugu Takasu,*,† Kimitsuna Watanabe,†,‡ and Gota Kawai*,[2]

Department of Industrial Chemistry, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba 275-0016; †Department of Chemistry and Biotechnology, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656; and ‡Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562

Received May 18, 2002; accepted May 20, 2002

A computational system, CSNA, for classifying RNA structures according to structural characters was developed. CSNA lists up all the hydrogen bonds and base-base stackings in the structures, and classifies the structures into sub-groups based on their patterns as the first step grouping. The frequency of each hydrogen bond or base-base stacking is calculated, the frequency score being defined as the sum of the frequency of existing hydrogen bonds or base-base stackings for each sub-group. Finally, the sub-groups are further classified into groups based on the frequency score defined in this study and the difference between the patterns. According to the frequency score, CSNA suggests a group that shares most frequently appearing hydrogen bonds and base-base stackings. CSNA was applied to the classification of the results of two individual simulated annealing calculations based on NMR information. It was found that CSNA could extract structures with lower energy without checking any energy term and could provide well converged groups as the lowest energy structures. Thus, CSNA could be a new tool for structural determination of nucleic acids.

Key words: base-base stacking, classification, hydrogen bond, NMR, RNA structure, structure determination.

The functions of bio-macromolecules, such as proteins and nucleic acids, depend on their 3D structures. Thus, determination or modeling of a 3D structure is important for clarifying the molecular mechanism of a function. Although X-ray crystallographic analysis and nuclear magnetic resonance (NMR) analysis are performed to obtain structural information at atomic resolution, these methods are not always applicable for larger biomolecules, especially for RNA due to its intrinsic flexibility and dynamics. Thus, computational methods are important for structural determination of such RNA.

As for structural determination by NMR, usually a hundred structures are calculated to obtain a set of converged results. However, due to the lack of information or intrinsic flexibility, the results of structural calculation for RNA molecules do not always converge well. In some cases, it is hard to know what happened by just checking the energy and NMR violations. On the other hand, it may be that even though the overall structures are not converged, some parts are converged locally, or more than one converged structure is simultaneously obtained. Thus, the classification of structures according to their similarities will be useful for analyzing the character of a set of NMR structures.

We developed a system for classifying a set of RNA structures based on the similarity of hydrogen bond and base-base stacking patterns. The preliminary version of the system has been applied to the results of structural modeling with the program MC-SYM (*1*). MC-SYM generates possible conformers of RNA with a given sequence by using the internal nucleotide structure database. 279 conformers produced by MC-SYM of 15-mer RNA hairpin were classified into 89 and 36 groups as to hydrogen bond and base-base stacking, respectively (*2, 3*). Even for such small molecules, there are so many hydrogen bond and base-base stacking patterns to obtain so many groups. Thus, in the present study, we introduced a second step.

We developed a system of computer programs, CSNA, standing for "Classification System for Nucleic Acid structure determination," to extract the hydrogen bond or base-base stacking information, and group structures in two steps according to the information. The first step is similar to the one we used previously (*2, 3*), and the second step was introduced in the present study. An outline of grouping is shown in Fig. 1. The system was applied to two individual sets of NMR structures: a 29-mer RNA, HE6 (*4*; PDB ID: 1L1W; Fig. 2A), corresponding to helix 6 of the human SRP RNA, and a 22-mer RNA, GBS/ωG (*5*; PDB ID: 1K2G; Fig. 2B), corresponding to the guanosine binding site and

[1] This work was supported by Grants-in-Aid for Scientific Research on Priority Areas (09278206) and High Technology Research from the Ministry of Education, Science, Sports and Culture of Japan, and the "Research for the Future" Program (JSPS-RFTF97L00503) from the Japan Society for the Promotion of Science.

[2] To whom correspondence should be addressed. Tel/Fax: +81-47-478-0425, E-mail: gkawai@ic.it-chiba.ac.jp

Abbreviations: ωG, 3′-terminal guanosine of the *Tetrahymena* group I intron; GBS, guanosine binding site of the *Tetrahymena* group I intron; GBS/ωG, 22-mer RNA corresponding to GBS and ωG; HE6, 29-mer RNA corresponding to helix 6 of human SRP RNA; NMR, nuclear magnetic resonance; rmsd, root mean square deviation; RNA, ribonucleic acid; SRP, signal recognition particle.

3′-terminal guanosine of the *Tetrahymena* group I intron (*6*). For both RNAs, NMR structures have been de-termined by the conventional method.

THEORY

Recognition of Hydrogen Bonds and Base-Base Stackings—For all pairs of acceptor and hydrogen atoms that are possibly included in hydrogen bonds, the distances are calculated, and any pair having a distance shorter than the pre-defined threshold (2.5 Å) is recognized as a pair forming a hydrogen bond. For the current system, distance is the only criterion for hydrogen bond discrimination.

To extract the base-base stacking, position and direction are defined for each base moiety. Position is defined as the average position of all but hydrogen atoms in the base. Direction is defined as the normal vector of the base plane. A base-base stacking pair is defined as a pair of bases for which the distance between the positions is less than 4.0 Å, and the angle between the directions is less than 30 degrees or more than 150 degrees.

The threshold values used for this detection were checked by using the well-known tRNA structure (data not shown).

First Step Grouping—The process starts with listing up all the hydrogen bonds and base-base stackings for each structure according to the criteria described above, and a "complete list" of hydrogen bonds and base-base stacking found in the set of structures is generated. Then, according to the complete list, a "bit list" for each structure is generated: the n-th bit in the bit list indicates the presence (1) or absence (0) of the n-th hydrogen bond or base-base stacking in the complete list. Finally, the structures are grouped by comparing the bit lists: each sub-group consists of structures having the same bit list.

Scoring the Sub-Groups—For each hydrogen bond or base-base stacking, the frequency of appearance in the set of structures is determined, and a "frequency list," in which the n-th number indicates the frequency of the appearance of the n-th hydrogen bond or base-base stacking in the set of structures, is produced. A "frequency score," which is defined as the inner product of the frequency list and the bit list, is calculated for each sub-group. This score represents the generality of the sub-group in the set of structures.

Second Step Grouping—To bundle the sub-groups sharing the frequently appearing hydrogen bonds and base-base stackings, weighted differences of structures for each pair of sub-groups are calculated as follows. A "difference list" is calculated as the logical exclusive or (XOR) between the bit lists. Then, a "structural distance" is defined as the inner product of the difference and frequency lists. The sub-group having the best (highest) frequency score is chosen as a reference sub-group of the best group. Then, the best group is defined as a group including the reference sub-group and sub-groups that exhibits the shorter structural distance from the reference group than a threshold. The next group is defined using another reference group that has the best frequency score among the remaining groups, and so on.

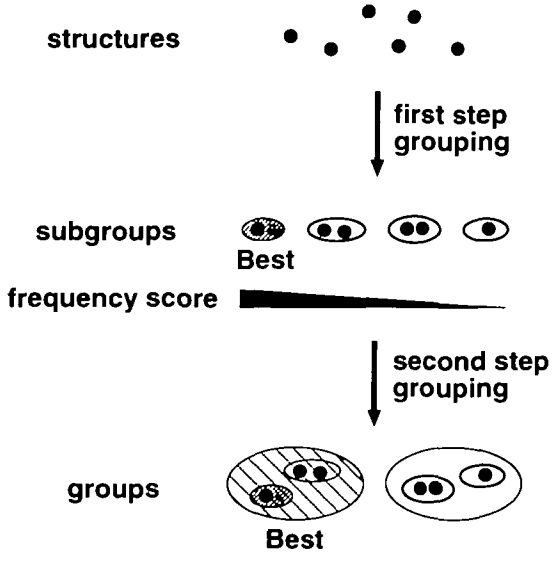

Fig. 1. **Strategy for classification.** The input for the system is a set of conformers of a molecule in the PDB format. According to the patterns of detected structural elements, the structures are classified into sub-groups. For second step grouping, the frequency score and structural distance are calculated. The sub-group that has the best frequency score is chosen as the reference sub-group of the best group. If the structural distance of each sub-group from the reference sub-group is less than a pre-defined threshold, the sub-group is added to the best group. The next group is defined using another reference group that has the best frequency score among the remaining groups, and so on.

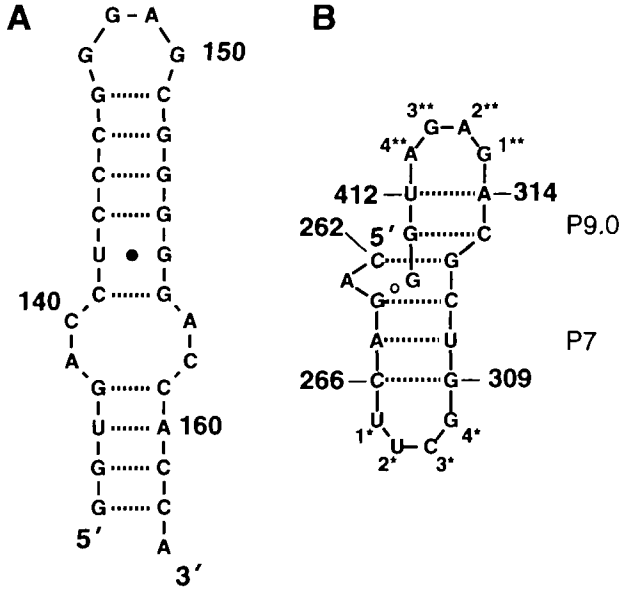

Fig. 2. **Secondary structures of RNAs.** (A) HE6 is a 29-mer RNA corresponding to helix 6 of human SRP RNA (*4*). The NMR structures have been obtained with the program Discover with the amber force field. (B) GBS/ωG is a 22-mer RNA corresponding to the P7 and P9.0 stems connected by two tetra loops (*5*). The NMR structures have been obtained by the program X-PLOR (*7*). Broken lines represent Watson-Crick base pairs. The filled circle represents a G-U wobble base pair. The open circle represents the interaction between the 3′-terminal guanosine (ωG) and guanosine binding site (GBS) of the *Tetrahymena* group I intron. Original numbering systems were used.

METHODS

CSNA was applied to the classification of 100 structures of HE6 and 400 structures of GBS/ωG, which had been obtained by structural calculations with Discover (Accelrys) and X-PLOR (*7*), respectively. For the first step grouping, each set of structures was classified into sub-groups, and as the second step grouping, each set of sub-groups was further classified into groups for each of several thresholds at the second step. In the case of GBS/ωG, the average structure and average root mean square deviation (rmsd) are calculated for four sets; all 400 structures, the 10 lowest energy structures, and the best groups with thresholds of 3 and 4. The average rmsd of a set was defined as the average rmsd between the average structure of the set and each structure of the set. The program CARNAL in AMBER (*8*) was used to calculate average structure and rmsd between structures.

RESULTS

Classification of 100 Structures of HE6—The new classification system, CSNA, was applied to a set of NMR structures of HE6, which consisted of both well-defined structures and less-defined ones. At the first step, 501 hydrogen bonds and 48 base-base stackings were detected, the resulting length of the bit list being 549. As a result of such a long bit list, 100 structures were classified into 100 sub-groups. In the next step, the 100 sub-groups were classified according to the structural distance (which is defined in the theory section). Figure 3A shows the relation between the threshold of the structural distance and the number of resulting groups for HE6. For grouping with both hydrogen bond and base-base stacking information (Fig. 3A, circles), if the threshold was less than 6, all sub-groups were classified into individual groups and each group consisted of a single sub-group, indicating that the smallest structural distance between the high score sub-group and other sub-groups is more than 6. On the other hand, if the threshold was larger than 20, the sub-groups were classified into a few groups. Figure 3A also shows the results of classification with only the hydrogen bonds (triangles) or base-base stackings (squares). Because the number of hydrogen bonds is 10 times greater than that of base-base stacking in the bit list, the smallest structural distance depends on the hydrogen bond patterns. However, with thresholds of more than 6, grouping depends on both the hydrogen bonds and base-base stacking. With the thresholds of 10 and 15, the 100 structures were classified into 75 and 28 groups, the best groups consisting of 7 and 41 structures, respectively. Figure 4, A and B, shows the number of structures in each group with the number of low-energy structures in each group. The total energy includes covalent (bond, angle) and non-covalent (van-der-Waals, electrostatic) energies, and improper and NMR-derived restraints terms were used as the calculated energy. It was shown that structures with the lowest energies were almost all classified into the top one or few groups.

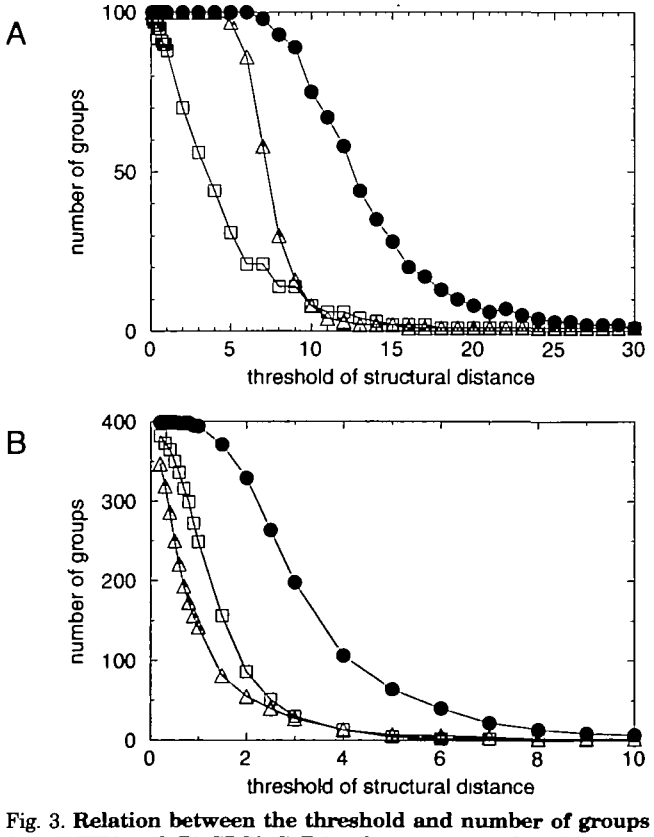

Fig. 3. **Relation between the threshold and number of groups for (A) HE6 and (B) GBS/ωG.** Triangles, squares, and circles represent hydrogen bonds, stacking, and both.

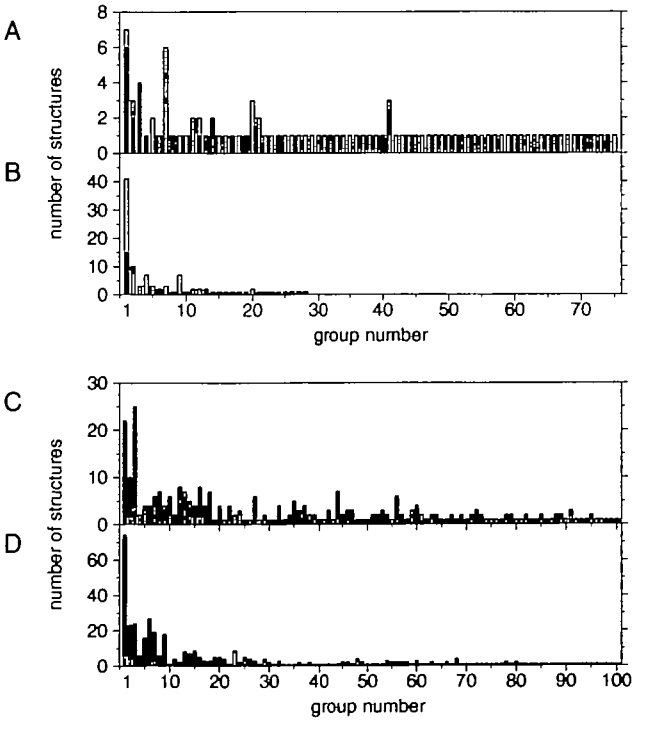

Fig. 4. **Distribution of the number of structures for each group.** Group number 1 represents the best group. The numbers of low energy structures are indicated by filled bars. (A) Results for HE6, with the threshold of 10. (B) Results for HE6, with the threshold of 15. (C) Results for GBS/ωG with the threshold of 3. (D) Results for GBS/ωG, with the threshold of 4. For panels C and D, only the top 100 groups are shown.

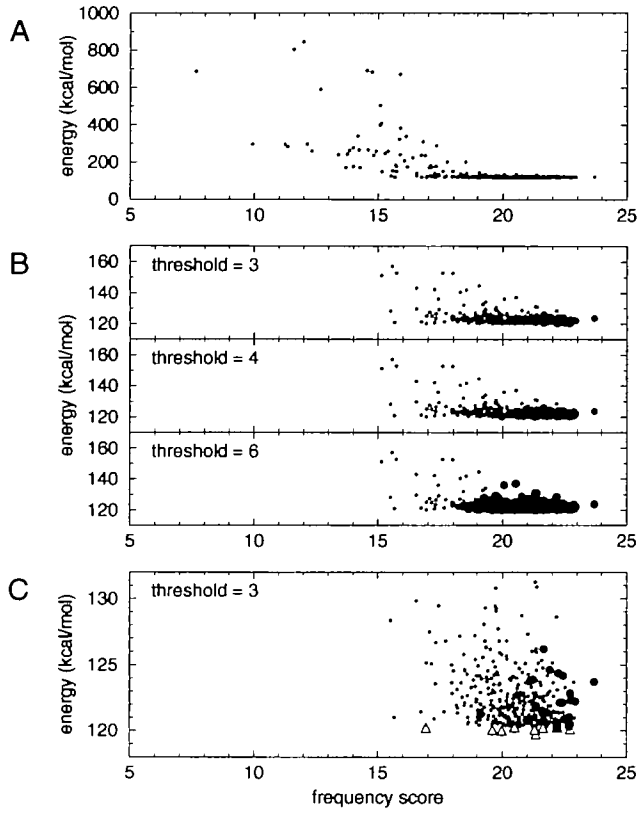

Fig. 5. **Relation between the energy and frequency score for the set of GBS/ωG structures.** The relation for all structures is shown in panel A. The filled circles in panels B and C represent the structures included in the best group. The triangles in panel C represents the 10 lowest energy structures.

Classification of 400 Structures of GBS/ωG—CSNA was also applied to a set of NMR structures of GBS/ωG. 200 hydrogen bonds and 37 base-base stackings were detected, the resulting length of the bit list being 237. As in the case of HE6, the 400 structures were classified into 398 subgroups. Figure 3B shows the relation between the threshold of the structural distance and the number of resulting groups for GBS/ωG. Compared to HE6 (Fig. 3A), the GBS/ωG structures could be grouped with a smaller threshold value, indicating that the GBS/ωG structures are more similar to each other than the HE6 ones. Figure 3B also shows that hydrogen bonds and base-base stackings contributed almost equally to the classification. Figure 4, C and D, shows the numbers of structures and low-energy structures for each group of GBS/ωG. The total energy includes covalent (bond, angle) and non-covalent (van-der-Waals) energies, and improper and NMR-derived restraints terms were used as the calculated energy. In these cases, the members of the best group and low energy structures did not match well.

Analysis of the Groups of GBS/ωG Structures—The relation between the potential energy and the frequency score was analyzed for the GBS/ωG structures, as shown in Fig. 5. Figure 5A shows a clear relation: lower energy structures have higher frequency scores. Judging from the energy, these structures were well converged and most of the structures have lower energy. In fact, 81% of the 400 structures exhibit no violations for distances and dihedral angles derived from NMR data.

TABLE I. **Average rmsd of some sets of structures of GBS/ωG.** rmsd between the average structure and each structure of a set were used to calculate average rmsd.

Set of structures	Number of structures	Average rmsd (Å)
All	400	3.56
Low energy structures	10	3.00
Best group with threshold of 3	22	2.92
Best group with threshold of 4	74	2.92

Figure 5B shows the distribution of the structures belonging to the best group for the thresholds of 3, 4, and 6. It was found that the best group for the threshold of 6 was almost the same with the lower energy structures, and CSNA could further classify the lower energy structures with smaller threshold values. Figure 5C shows that the members of the best groups with the threshold of 3 were concentrated in the highest score region and did not overlap the 10 lowest energy structures (triangles).

DISCUSSION

Previously, we developed a classification system that judges structural similarity according to two randomly selected reference structures (*2*). With that system, the results are biased by the reference structures. Then, the algorithm was improved (*3*) and a new system, CSNA, was proposed, as described above. The superiority of this new algorithm is the independence of reference structures and adjustability of the grouping threshold. Figure 3 shows that both the hydrogen bond and base-base stacking contribute to the grouping, indicating that detection of hydrogen bonds and base-base stacking works, and suggesting that hydrogen bonds and base-base stacking are similarly important for RNA structure. When the threshold of 10 was used in the case of HE6, the members of the best four groups were almost the same with the lowest energy structures (Fig. 4A). This also indicates the validity of this system. In the case of GBS/ωG, the members of the best group were concentrated in the highest score region and not located in the lowest energy region (Fig. 5C). However, if the threshold of 6 was used, the members of the best group, which consisted of 275 structures, were almost the same with the lower energy structures (Fig. 5B, bottom panel).

Figures 4A and B, and 5A suggest that the structures having higher frequency scores are more structurally stable. Simply, the structures have more hydrogen bond and base-base stacking interactions, indicating that the structures have lower energy. Table I shows the average rmsd for the set of low energy structures and the best groups for GBS/ωG with the thresholds of 3 and 4. It was found that both average rmsd values for the best groups are lower than the set of the lower energy structures, suggesting an alternative way of selecting "converged" structures from a set of calculated structures. It should be noted that the best group chosen by CSNA is not always a feasible structure and, also, CSNA may provide more than one best group. CSNA can analyze the characteristic of the structure set, and suggest groups sharing most frequently appearing hydrogen bonds and base-base stackings.

There are several classification systems for structures, such as FSSP (*9*), CATH (*10, 11*), and SCOP (*12*) for protein structures, and SCOR (*13*) for RNA structures. Furthermore, an automated procedure for assigning CATH and SCOP classifications to proteins whose FSSP scores are available has been reported (*14*). However, the targets of these systems differ from that of CSNA, which is a set of conformers of one molecule. On the other hand, the program NMRCLUST (*15*) clusters a set of structures of one molecule. But it cannot suggest the best group because it clusters based on the pair wised rmsd, which does not reflect molecular stability. More importantly, if local structures in a set of structures converge, CSNA must be able to detect it because the system classifies structures according to their local structural information, such as hydrogen bonds and base-base stacking. In other words, CSNA may be used as a tool for analyzing the local structure of RNA. On the other hand, the local structures of proteins can be analyzed by using a C^{α}–C^{α} distance map (*16*), which reflects local structures.

In spite of the simplicity of the algorithm used in the classification system CSNA, it could extract a set of lower energy or well-defined structures. Thus, CSNA is a rapid and useful way of analyzing a set of structures. In particular, it will be useful for classifying not well-converged structures to extract common features of a set of structures or molecular parts, which are converged well. The authors are preparing to provide CSNA function in a Web site.

The authors are grateful to Prof. S. Yokoyama for the use of the NMR structures, and Drs. A. Kitamura and T. Sakamoto for the helpful discussion.

REFERENCES

1. Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E., and Cedergren, R. (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* **253**, 1255–1260
2. Takasu, A., Kawai, G., and Watanabe, K. (1999) Development of a system to classify 3D structural character of RNA. *Nucleic Acids Symp. Ser.* **42**, 233–234
3. Takasu, A., Kawai, G., and Watanabe, K. (2000) Classification of 3D structural character of RNA by hydrogen bond and base stacking. *Nucleic Acids Symp. Ser.* **44**, 227–228
4. Sakamoto, T., Morita, S., Tabata, K., Nakamura, K., and Kawai, G. (2002) Solution structure of SRP19 binding domain in human SRP RNA. *J. Biochem.* **132**, 177–182
5. Kitamura, A., Muto, Y., Watanabe, S., Kim, I., Ito, T., Nishiya, Y., Sakamoto, K., Ohtsuki, T., Kawai, G., Watanabe, K., Hosono, K., Takaku, H., Katoh, E., Yamazaki, T., Inoue, T., and Yokoyama, S. (2002) Solution structure of an RNA fragment with the P7/P9.0 region and the 3′-terminal guanosine of the *Tetrahymena* group I intron. *RNA* **8**, 440–451
6. Michel, F., Hanna, M., Green, R., Bartel, D.P., and Szostak, J.W. (1989) The guanosine binding site of the *Tetrahymena* ribozyme. *Nature* **342**, 391–395
7. Brunger, A.T. (1992) *X-PLOR, Version 3.1. A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT
8. Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., I, Ross, W.S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V., Cheng, A.L., Vincent, J.J., Crowley, M., Ferguson, D.M., Radmer, R.J., Seibel, G.L., Singh, U.C., Weiner P.K., and Kollman, P.A. (1997) *AMBER 5*, University of California, San Francisco
9. Holm, L. and Sander, C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* **22**, 3600–3609
10. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH—A hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108
11. Pearl, F.M.G., Lee, D., Bray, J.E, Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., and Orengo, C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28**, 277–282
12. Conte, L.L., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**, 264–267
13. Klosterman, P.S., Tamura, M., Holbrook, S.R., and Brenner, S.E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.* **30**, 392–394
14. Getz, G., Vendruscolo, M., Sachs, D., and Domany, E. (2002) Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins* **46**, 405–415
15. Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **9**, 1063–1065
16. Carugo, O. and Pongor, S. (2002) Protein fold similarity estimated by a probabilistic approach based on C^{α}-C^{α} distance comparison. *J. Mol. Biol.* **315**, 887–898