3514–3522 *Nucleic Acids Research, 1997, Vol. 25, No. 17*

ESSA: an integrated and interactive computer tool for analysing RNA secondary structure

F. Chetouani, P. Monestié[1], P. Thébault, C. Gaspin[1] and B. Michot*

Laboratoire de Biologie Moléculaire Eucaryote du C.N.R.S., Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France and [1]Station de Biométrie et d'Intelligence Artificielle, I.N.R.A., Chemin de Borde-Rouge, Auzeville BP 27, 31326 Castanet-Tolosan Cedex, France

Received April 17, 1997; Revised and Accepted July 7, 1997

ABSTRACT

With ESSA, we propose an approach of RNA secondary structure analysis based on extensive viewing within a friendly graphical interface. This computer program is organized around the display of folding models produced by two complementary methods suitable to draw long RNA molecules. Any feature of interest can be managed directly on the display and highlighted by a rich combination of colours and symbols with emphasis given to structural probe accessibilities. ESSA also includes a word searching procedure allowing easy visual identification of structural features even complex and degenerated. Analysis functions make it possible to calculate the thermodynamic stability of any part of a folding using several models and compare homologous aligned RNA both in primary and secondary structure. The predictive capacities of ESSA which brings together the experimental, thermodynamic and comparative methods, are increased by coupling it with a program dedicated to RNA folding prediction based on constraints management and propagation. The potentialities of ESSA are illustrated by the identification of a possible tertiary motif in the LSU rRNA and the visualization of a pseudoknot in S15 mRNA.

INTRODUCTION

Understanding of RNA structure–function relationships requires complex working strategies which make large use of computer programs in addition to bench experiments. They involve structural and functional predictions deduced from sequence analysis and interpreted in the light of a wide and diverse knowledge provided in part by computer approaches. The determination of the 3D folding of Group I ribozymes is probably the best example of structured RNA analysis strategy (1). The first step consists of the prediction of reliable secondary structure folding which can be defined as the set of C:G, A:U Watson–Crick and G:U wobble pairs allowing their readable representation in two dimensions. Three approaches were developed which have in common the knowledge of the primary structure, with a view to identify the set of hydrogen-bonded nucleotides involved in stabilizing their folding. The measure of nucleotide accessibility to chemical and enzymatic structure probes allows the identification of paired and unpaired positions (2,3), thermodynamic optimization proposes optimal and suboptimal foldings (4), whereas comparative analysis consists of a systematic search for compensatory mutations in an alignment of homologous RNA sequences from several organisms (5–7). These three approaches provide different structural information and have their own limitations, but the third one has the main advantage of pointing directly to biologically significant structural features. Nevertheless, the determination of the secondary structure folding often requires the conjunction of these three complementary methods. Thus, structural information from different origins must be used simultaneously in order to converge towards a model that is in agreement with all available data. This has led to the consideration of RNA modelling as a constraint satisfaction problem (8,9).

The second step is the representation of these secondary structure folding models which must be comprehensive enough to serve as a support for the evaluation, refinement and management of folded RNA, but also for their own interpretation with a view to predicting new structure–function relationships according to the user knowledge. Further analyses include comparison of models, searching for functional structural motifs, checking for pseudoknots and possible alternative interactions, identification of co-varying positions and higher order interaction. Interpretation of results necessitates the integration of many different types of information encompassing all those related to the structure and the function of the molecule studied, or from homologous molecules. Among them, accessibility to enzymatic and chemical probes, position and type of modified nucleotides, RNA–RNA and RNA–protein inter-molecular contacts are essential. The extensive viewing and management of this knowledge is a prerequisite to point to those structural features which could play biological roles. This reveals the necessity of possessing highly interactive and integrated computer programs which allow an easy representation of RNA secondary structures and the direct investigation of the model produced through a set of diversified analysis functions.

Several programs are available, each dedicated to one of the numerous problems posed by RNA secondary structure modelling. Thus, softwares predict RNA folding or probabilities of pairing from individual sequences with thermodynamic rules (10–12). Others focus on the representation and display of secondary structure (13,14) or the identification of co-varying positions, Watson–Crick or not, in aligned sequence datafiles (15–17). The

*To whom correspondence should be addressed. Tel: +33 5 61 33 58 65; Fax: +33 5 61 33 58 86; Email: bmichot@ibcg.biotoul.fr

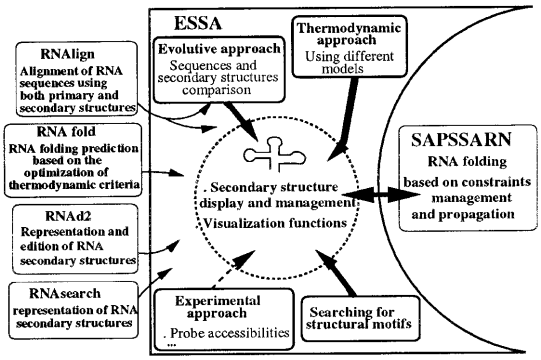


Figure 1. ESSA relationships with RNA strategy analysis. Within ESSA program (boxed by a thick full line), the type of communication between the secondary structure vizualizing-editing core (circled with a dotted line) and the main functions (boxed by a full line) is depicted by different arrows: dotted arrows correspond to keyboard input whereas thick simple arrows indicate a full integration of functions. Inter-relations with other programs are denoted by simple and thin arrows for file input while the thick and double arrow points to a full communication protocol developed with SAPSSARN.

management and analysis of specialized and aligned sequence files such as those containing ribosomal RNAs have also led to the development of specific softwares (18,19). Unfortunately each of these programs has its own application field and works independently, usually under different operating systems, when they should be closely linked with a view to fitting with the RNA analysis. Softwares rarely include several of these programs. The GCG package (20) connects thermodynamic secondary structure prediction programs with several modes of representation, whereas a more recent program (21) combines viewing of the structure and the use of probabilities of pairings.

The aim of ESSA is to propose an interactive approach of RNA secondary structure analysis which integrates their representation, the visualization of various types of information, and analysis with a view to covering the most important aspects devoted to a better understanding of their structure–function relationships. ESSA was also designed to communicate with other programs dedicated to RNA secondary structure. More particularly a communication protocol was developed with SAPSSARN (9), a program which relies on the probabilities of pairing (11) associated with constraints management and propagation to help with the prediction of RNA folding. We present here the numerous applications of ESSA and give two examples of its predictive possibilities. The first concerns a search in the LSU rRNA for a structural motif involved in a tertiary contact first identified in autocatalytic group I introns (22). The second illustrates its capacity to fold RNA and predict complex interactions owing to ESSA–SAPSSARN communication. We show how a pseudoknot involving a conformational switch in the mRNA of ribosomal protein S15 (23) can be viewed.

MATERIALS AND METHODS

Various import formats are supported by ESSA including RNAsearch (13), RNAd2 (14), FoldRNA of the GCG package (20) and RNAlign (24). More specifically, ESSA exchanges data with SAPSSARN through a communication protocol based on a client server model. The X server acts as an intermediary between ESSA and SAPSSARN client applications. Data exchanged are

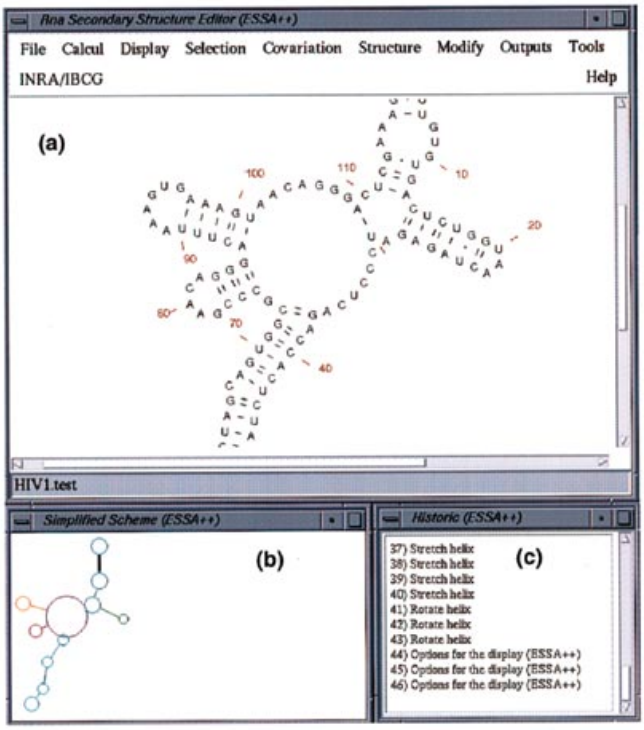


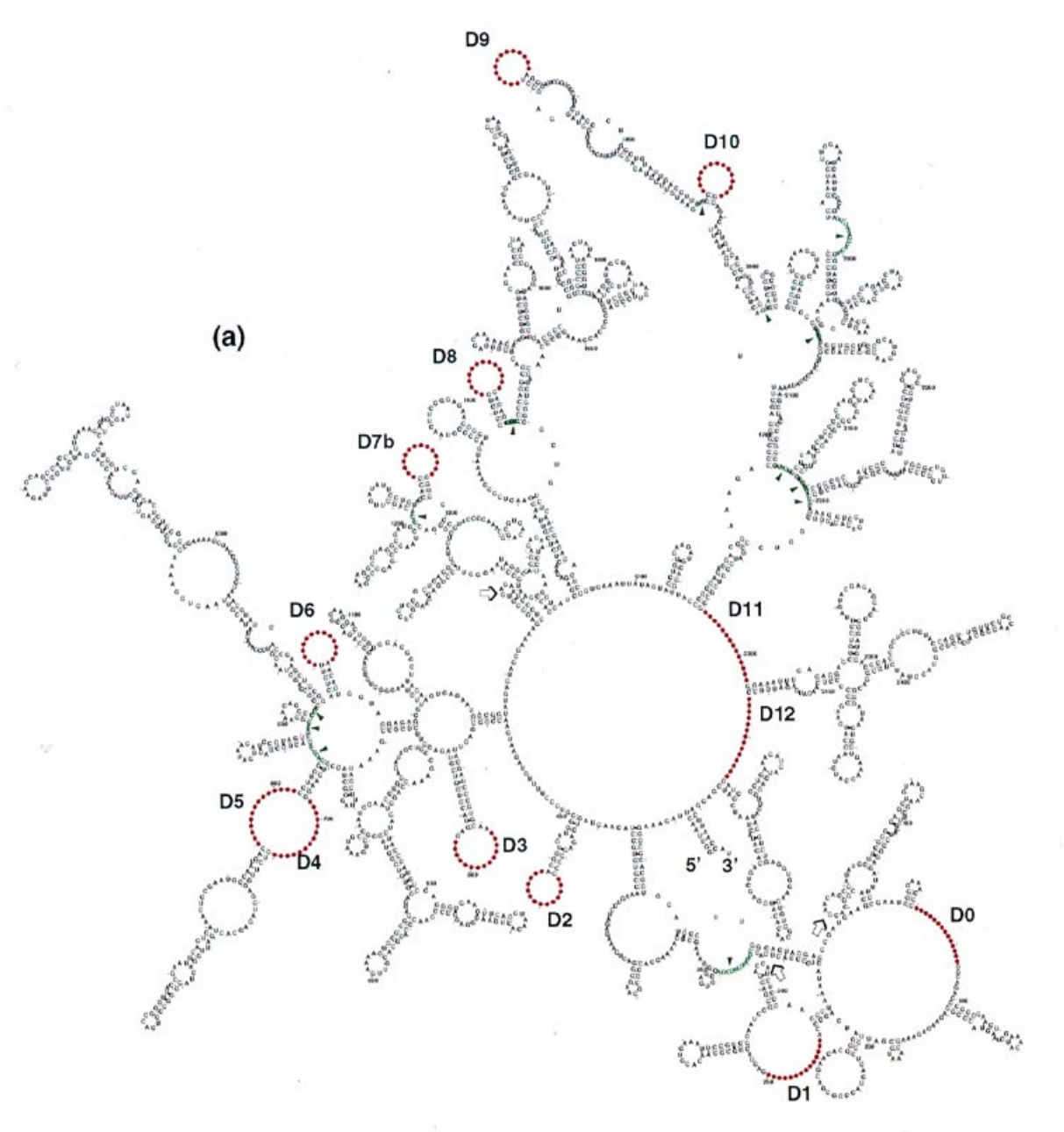
Figure 2. The three windows of ESSA interface. The main window (**a**) displays a graphical representation of a secondary structure in the drawing area. Editing and analyses are applied on structural elements of the secondary structure thanks to options offered in the menu (upper part of the window) and by mouse-click within the drawing windows. A backbone view (**b**) is also displayed where each loop is represented by a circle and each stem is represented by an edge. A third window (**c**) records informations relative to selected objects and functions applied on selections during a session.

the sequences and the secondary structure: once both applications have run and the sequence is known, each modification of the secondary structure in any application is recorded as a request to the other and is managed by the server.

A postscript output of the drawing can be generated, then saved in a file or printed directly. Outputs include all the results of editing and analysis functions which are displayed on secondary structure models by coloring, numbering, changing the size of nucleotides and symbols. All the labelling can also be saved for further analyses. ESSA is written in C ANSI language and runs on SiliconGraphics (under Irix 5.3 or later), SUN (under Solaris 2.4 or later) and HP (under HP-UX 9.0 or later) Workstations within the X11/MOTIF environment. The program is self documented and a user manual is available. An e-mail address (essa@toulouse.inra.fr) is available for any request for an executable version but also to receive comments, suggestions and questions from users.

RESULTS

ESSA relies on the representation of RNA folding, a central core which proposes a first set of functions dedicated to the management of secondary structures (Fig. 1). This display serves also as a support to emphasize the viewing of remarkable structural features and to update diverse knowledge via the keyboard. Running on this core, a set of analysis functions was developed allowing search for structural motifs, calculation of the



thermodynamic stability of any substructure and folding comparisons. Moreover, ESSA communicates by files with other programs dedicated to RNA secondary structure. Thus, comparative analysis functions provided by our program are linked with specialized and structured databanks produced by RNAlign. ESSA is also connected with RNA secondary structure models computed by several other programs, among them the outputs from FoldRNA. Finally, we have developed a communication protocol owing to a real time exchange of information with SAPSSARN. This program integrates the thermodynamic approach with probabilities of pairing and propagation of constraints given by the user with a view to offering an effective tool in terms of structure prediction.

This rich integration of display and analysis functions was realized through an interface designed to be intuitive and easy to use. To achieve this aim, ESSA opens three windows, two of which are directly devoted to the working session. The main window (Fig. 2a) visualizes RNA folding models and all the information manageable by ESSA. This window is the essential way to communicate with the program by acting either directly on the display or through a menu proposed in its upper part. By contrast, the backbone view (Fig. 2b) contains only a simplified representation of RNA secondary structure in which stems are replaced by single straight lines and loops by polygons with different colours given for each branch rooted on a multibranched loop. Interactions through this window are essentially restricted to the management of structural features. These two windows are interconnected and each action in one of them is updated in real time in the other. The third window (Fig. 2c) records the successive actions performed during a working session.

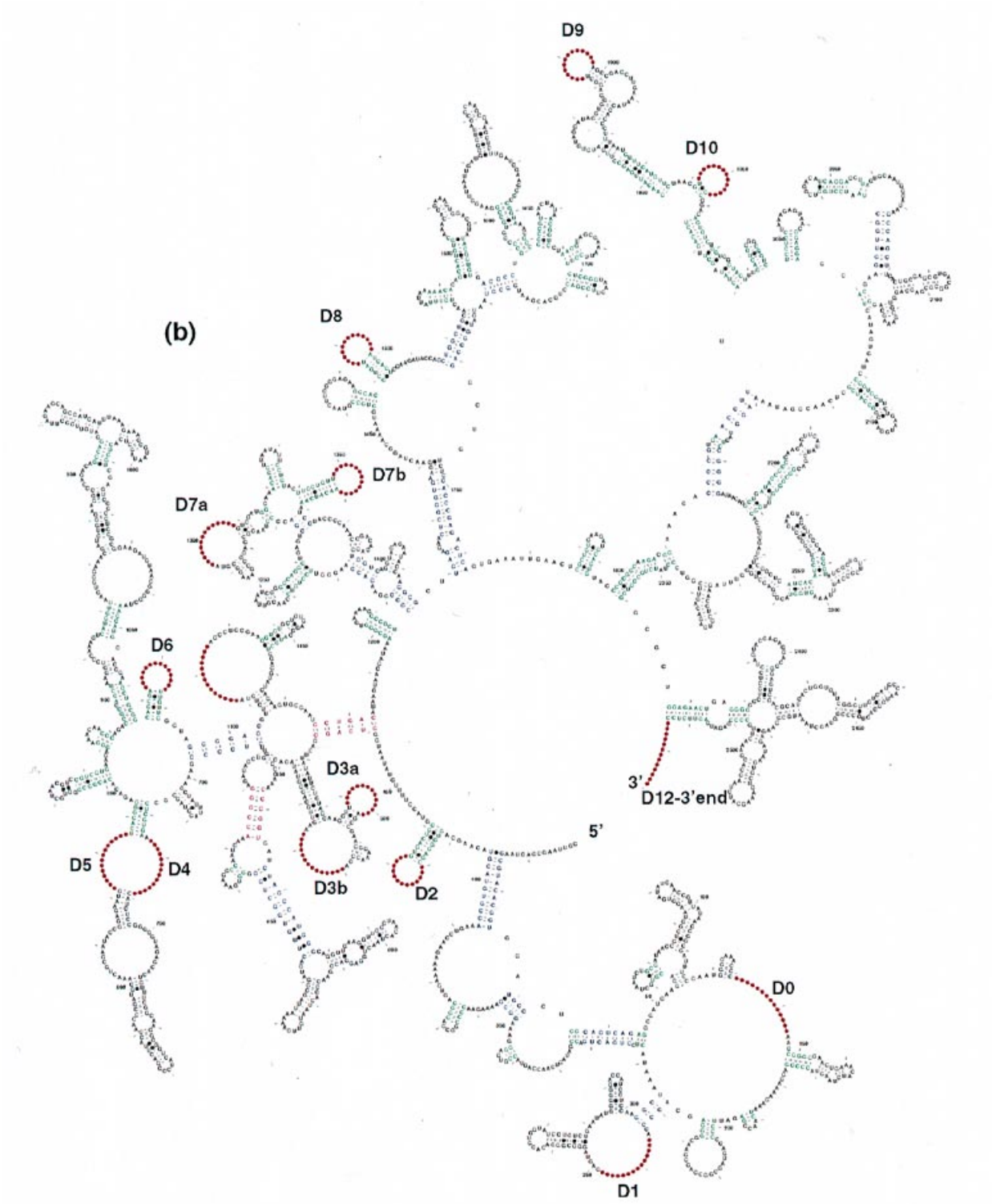


Figure 3. Comparison of the two modes for producing secondary structure drawings. The LSU rRNA conserved core of secondary structure was drawn (**a**) in the fully automatic mode for *Bacillus subtilis* and (**b**) in the interactive mode for *Escherichia coli*. In (a) green portions point to sequences tracts where nucleotides are overwritten, whereas arrows locate closely displayed structures. In (b), coloured stems show structural elements which were either rotated (green) or stretched (pink) or rotated and stretched (blue) in the initial polygonal display. Divergent domains are identified according to Hassouna *et al.* (25). D0 locates a short 23S variable region in eubacteria which corresponds to the internal spacer ITS2 in eukaryotes.

According to the aim of ESSA, the first prerequisite to every investigation is the display of RNA molecules. When only the sequence is available it is displayed as a circle, otherwise the secondary structure folding is drawn. We have chosen a polygonal representation which has the advantage of clear readability and is well adapted for diverse and comprehensive additional labelling.

Two programs were implemented which are complementary and offer a quick visualization of RNA secondary structures without overlap and the possibility to display homologous structures in a similar shape. One is fully automatic and uses a backtracking algorithm tuned by a set of parameters which makes it possible to adjust the deformations imposed on a strict polygonal display to avoid overlaps while preserving the compactness and the aesthetic character of the drawing (Fig. 3a). This method is very efficient and fine adjustment of parameters allows us to draw the secondary structure of molecules as long and complex as the universal core of secondary structure of the large ribosomal subunit RNA (LSU rRNA) which encompasses about 2500 nucleotides organized in 148 stems. Nevertheless, depending on the molecule, the parameter values required for a complete suppression of overlaps may introduce important distorsions. This can result in a loss of readability due to a close proximity between several substructures, and the overwriting of nucleotides in several internal loops. These two problems occur at 3 and 13 sites in the drawing of the LSU rRNA conserved core respectively. Therefore, a complementary approach is proposed (Fig. 3b) in which a strict polygonal drawing is first computed. Then, the relative organization of secondary structure features can be interactively managed by the user who chooses his own strategy to remove overlaps through a minimal number of interactions by using a set of editing functions. A first one identifies and progressively straightens the subdomains which are highly compacted by very dissymetric internal loops. This automatic tool is particularly useful when the display is obscured by numerous overlaps despite the colour code of the backbone view. Other editing functions are interactive and consist of rotation/displacement of any subdomain. They largely use the backbone display to make the identification and manipulation of structural features easier. This essentially interactive approach allows the organization of the folding of any RNA molecule according to the user's wishes. It allows easy drawing of molecules as long as the LSU rRNA universal core.

Once an RNA sequence or a secondary structure is displayed, a set of functions becomes available for fine labelling and deeper analysis of the molecule. Most of them can be used either on the entire molecule or on a selected set of any subregion making the current selection a crucial feature of ESSA. Facilities are given to select stems, helix regions or domains directly through a menu of basic elementary selections. The current selection appears red on the display. Interactive editing functions were developed to manage the base pair set and label the molecule with numerous and diverse types of information. Secondary structure models can be refined and updated according to new structural results using functions which pair/unpair two elementary selections. The pair function necessitates that (i) neither base of the elementary selection is already involved in a pairing, (ii) the two elementary selections have the same size and (iii) they do not create a pseudoknot. By contrast, pairings are not verified for Watson–Crick/wobble pairs allowing the user to introduce any non canonical pairing according to experimental or evolutionary data. These tools are suitable to create *de novo* a folding from the display of an RNA sequence. Rich labelling possibilities are also offered through several basic tools. A palette of 18 colours is proposed to highlight the diverse structural or functional informations available for a molecule (for examples see Figs 3 and 5). This can be enhanced by changing locally the mode of display. One can represent or not the pairing between bases, replace

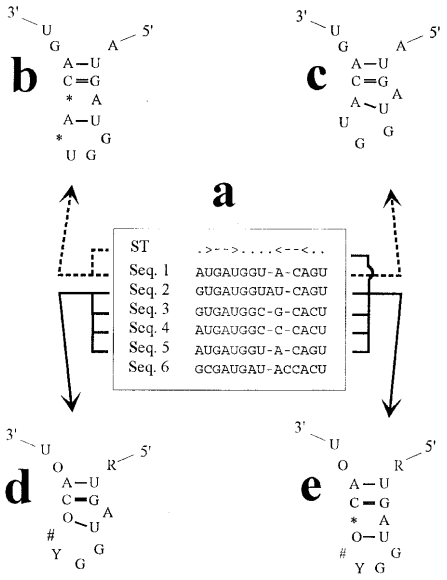


Figure 4. Production of consensus in sequence and secondary structure. (**a**) A hypothetical specialized and aligned databank is shown which contains six sequences identified from Seq. 1 to Seq. 6 where hyphens denote gaps. The line ST contains a codage specifying paired nucleotides according to (24). In (**b**) the folding of Seq. 1 is drawn according to the universal folding whereas in (**c**) the same sequence is drawn according to its own folding. In (**d**) and (**e**) a consensus in sequence of Seq. 2 to Seq. 5 is drawn respectively according to the specific folding which is common to this group of species and to the universal folding. In (b) and (e) stars denote gaps which are present in the sequence(s) displayed by reference to the universal folding. In (d) and (e), when a set of sequences are compared, '#' stand for gaps, 'Y' is for pyrimidines, 'R' for purines and 'O' for any base, according to the input value for consensus.

nucleotides by a dot, bind them with edges, display or not hydrogen bonds. ESSA also proposes a diversified set of symbols to label each nucleotide, among which several are especially targeted at the management of results coming from enzymatic and chemical structural probe accessibilities (Fig. 6).

The identification of nucleotides or sequence motifs to label is made easier by diverse numbering possibilities. In particular any value can be assigned to the first nucleotide of the molecule, even negative, with a view to adapting the numbering to any convention. We have also developed a word searching procedure to directly reach sequence motifs which can be degenerated. Resulting occurrences are inserted in the current selection and thus coloured on the model. Then, each occurrence can be conserved or removed according to its significance as evaluated by the user. Moreover, we have intimately coupled several aspects of the evolutionary and thermodynamic approaches. Concerning the evolutionary one, either all or a subset of sequences can be extracted from specialized databanks containing aligned sequences and an encoding of their secondary structure interactions. Then, four types of display can be produced depending on whether only one or several sequences are extracted and whether the secondary structure is designed as a universal consensus or is specific to the extraction (Fig. 4). The visualization of one sequence according to its specific folding will remove all the gaps from the display, whereas when a set of sequences is selected only the gaps which are common to all of them will be removed. When sequences are drawn according to the universal consensus, all the gaps inserted in the alignment are displayed and each drawing will have exactly

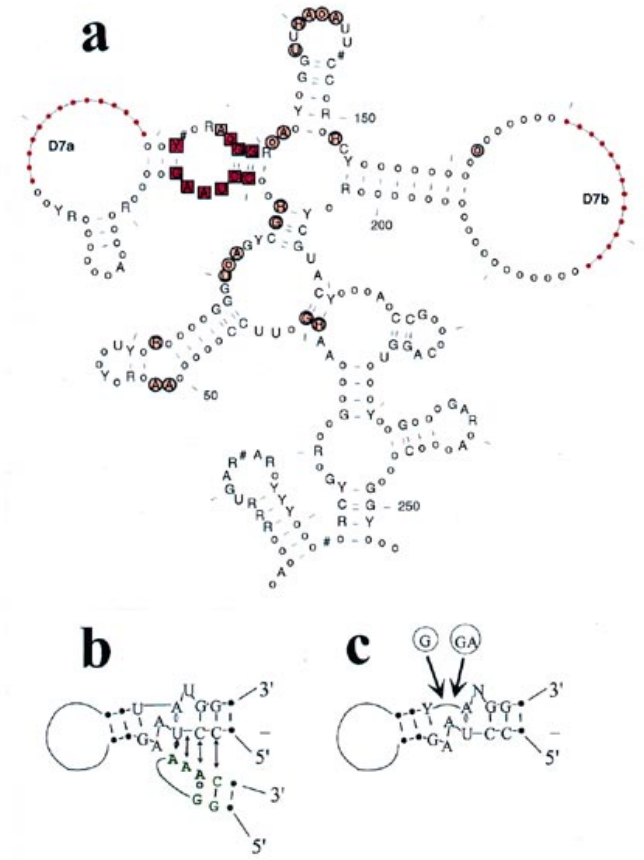


Figure 5. Consensus of secondary structure for the LSU rRNA region encompassing a group I-like tertiary motif. (**a**) Fifty sequences chosen in the three major kingdoms (eubacteria, eukaryotes and archaebacteria) were selected to be representative of each major branching and a 90% consensus was produced. Divergent domains are replaced by red dotted lines. Nucleotides boxed by squares locate the group I intron-like motif. Positions protected from nucleases and chemical probes by ribosomal protein L23 (33,34) are coloured in yellow and circled, except the yellow-squared 'A' which belongs to the Group I-like tertiary motif. (**b**) Structure of the canonical Group I motif showing its interaction with a GAAA tetraloop (26). (**c**) Structure of the motif identified in the LSU rRNA molecule. The two arrows point to the insertion site of 1 or 2 nucleotides (circled) respectively in eubacteria and eukaryotes whereas in archaebacteria the situation is more complex with insertions of 1 or 2 nucleotides, exceptionally 3.

the same shape. Moreover, each time two or more sequences are selected, a consensus at each position is computed according to a given value. ESSA also offers the possibility of displaying a sequence from a file of aligned sequences on any drawing previously made from this file and according to its consensus folding. Thus all drawing produced from structured and aligned databanks will have exactly the same shape. In addition, this function allows the display of any information already entered for one sequence of a file, on any other sequence or set of sequences of the same file. As for the thermodynamic approach, ESSA calculates the energy of any substructure or set of substructures included in a selection. A default thermodynamic model is given with options as to whether to consider or not loops such as tetraloops, multibranched loops, or internal loops and bulges, but any other thermodynamic model could be used instead.

Finally, ESSA has predictive capabilities through communication with SAPSSARN based on facilities provided by X Window. This program implements a constraint satisfaction approach of 2D folding in order to explore efficiently alternative secondary structures on the basis of thermodynamic, experimental or phylogenetic data. The set of candidate pairs to a structure evolves according to propagation of constraints which come from any structural data and can be added or removed at any time of a working session. When both SAPSSARN and ESSA work on the same RNA secondary structure, they communicate each change in the base pair set. Thus, each (un)pairing in SAPSSARN induces ESSA to compute a new representation of the secondary structure, whereas, each (un)pairing in ESSA is propagated as a constraint in SAPSSARN, but only if this (un)pairing is allowed by SAPSSARN. Pseudoknots, which are used as constraints, when allowed in SAPSSARN become also allowed in ESSA where they appear as coloured bases.

Among the wide possibilities of investigation inherent in ESSA we chose to emphasize a search for a secondary structure motif in the light of the evolutionary approach and the communication with SAPSSARN through the identification of a pseudoknotted interaction. In the first example, we searched the LSU rRNA for the presence of the single structural motif involved in a 3D interaction which was identified in several RNA molecules. This motif consists of the 11 nucleotides 5′-(CC)UAA(G)... (U)AU(GG)-3′, base paired in a precise stem–loop configuration, which interacts with a GAAA tetraloop (Fig. 5b). We have first produced a consensus of the conserved core of secondary structure for the LSU rRNA (data not shown). Then we searched successively for the two sequence segments which constitute the 11 nt secondary structure motif. The 5′ and 3′ sequence segments were found respectively three and two times. The visual examination of the structural environment of each of them revealed only one significant homology with the searched motif in a region interrupted by non-conserved domains (Fig. 5a). This stem–loop structure is one of the best conserved motifs of this region. The (CC)UAA(G) sequence motif is strongly conserved in all species analysed. By contrast, the (U)AU(GG) sequence segment could not be identified by our searching function since it is split into two parts by an insertion of 1 or 2 nucleotides, exceptionally 3, according to the three main kingdoms (Fig. 5c). Nevertheless, it is remarkable that not only the secondary structure, but also the higher structural features which are essential for the organization of the spatial arrangement of the receptor site are preserved. Thus, the A–A platform, the A–U reverse Hoogsteen hydrogen bond which is involved in a triple base interaction with a nucleotide of the tetraloop and the two base-paired G which also bind the GAAA tetraloop are strongly conserved.

In a second example we used the communication protocol developed between ESSA and SAPSSARN to point to a pseudoknot structure in S15 mRNA. In a first step, the communication between SAPSSARN and ESSA was established in the case of the wild-type RNA sequence. In SAPSSARN pseudoknots were forbidden, a 3 nucleotide minimum stem size was imposed and a threshold of 5 was retained for possible pairs (Fig. 6a). In ESSA we labelled the sequence with probe reactivities. Then pairs were selected in ESSA to form the secondary structure. Once constraints of pairing were propagated, only one solution remained possible in SAPSSARN search phase (Fig. 6b). In a second step, the same protocol was used for the analysis of an S15

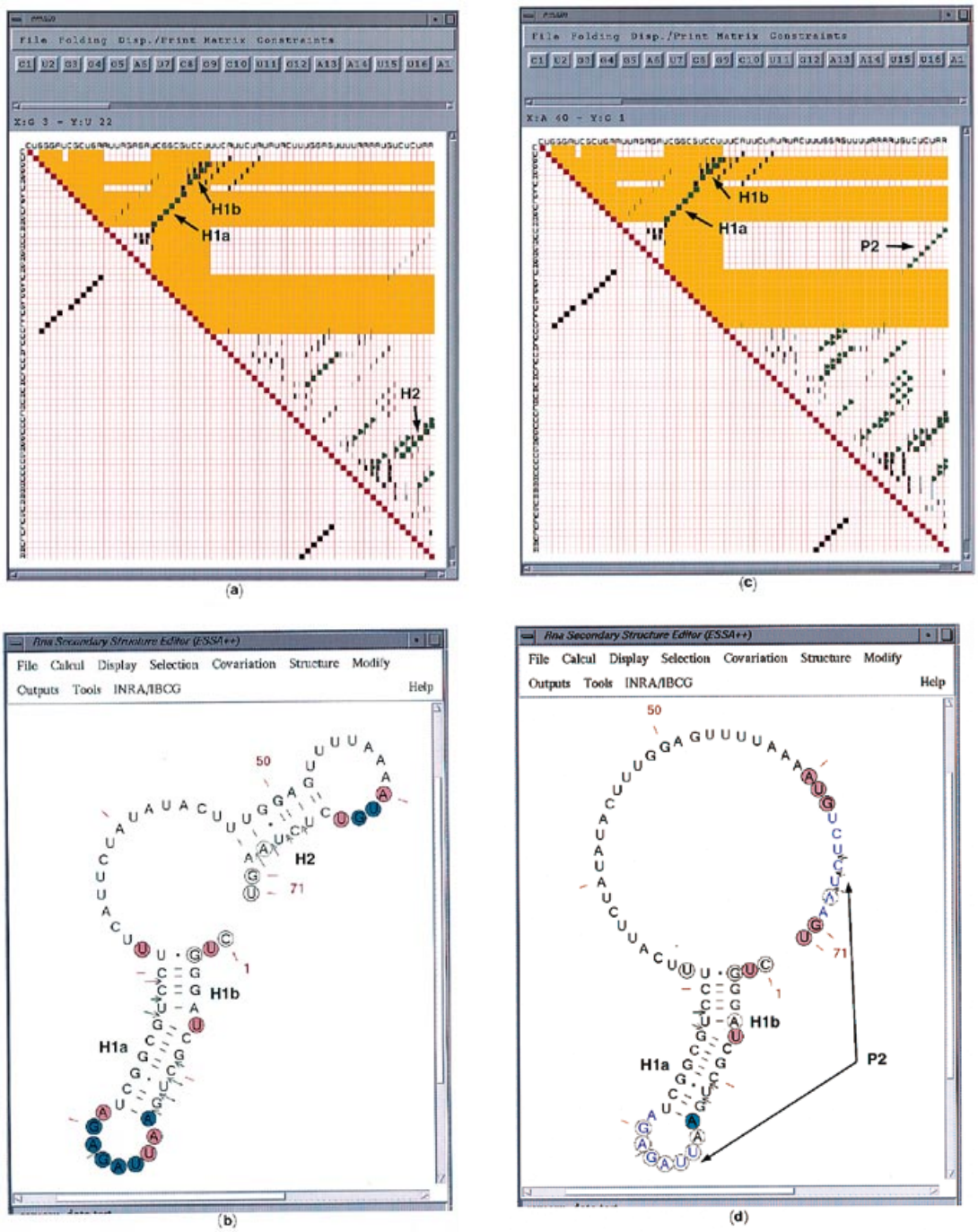


Figure 6. Secondary structures folding prediction in S15 mRNA. In (**a**) and (**c**), the upper half-matrices give the probabilities of pairing of each base with each other computed by SAPSSARN following (11). Black squares represent probabilities of pairing within a logarithmic scale from 0 to 9. Yellow squares point to bases already selected in pairs. Green circles represent candidate pairs to select after the propagation of constraints. The lower half-matrices give the optimal secondary structure (based only on thermodynamic data) represented by black squares. In (**b**) and (**d**) secondary structures of S15 mRNA are tagged on the basis of data from Philippe *et al*. (23). In both cases, the sequence of the wild type was represented and labelled by experimental data obtained in cases of the wild-type (b) and the CFP5517 mutant (d). Reactivity of Watson–Crick positions were obtained using DMS (A,C), CMT (U) and kethoxal (G). Cuts were realized using the RNase V1. Reactivities to chemical probes are represented with blue circles (strong reactivity), pink circles (medium reactivity), white circles (moderate reactivity) or dashed line circles (marginal reactivity). RNase V1 cuts are indicated by arrows the size of which is proportional to the intensity of the cleavage. In the case of the wild type S15 mRNA (b), possibilities of pairing displayed in the corresponding matrix (a), are visualized after selecting H1a, H1b and H2 which are those which better fit probe accessibilities. In contrast, in the case of the CFP5517 mutant (d), stem H2 does not appear as the optimal folding according to the probe accessibilities in H1a hairpin loop (d).

mRNA mutant (Fig. 6c and d). Remarkably in this case, chemical probe accessibilities differ from the wild type RNA. The low reactivity of H1a hairpin loop to Watson–Crick chemical probes reveals its involvement in Watson–Crick base pairs pointing to a probable pseudoknot interaction. At that time, we removed in SAPSSARN the constraint on the absence of pseudoknot with a

threshold fixed to 1. There appears then in SAPSSARN's matrix a new candidate stem called P2 which is incompatible with the stem H2 since its 3′ strand is involved in a pairing with the H1 hairpin loop. This new pairing was selected in SAPSSARN's matrix and communicated to ESSA leading to the coloration of the pseudoknot P2, visualizing in this way the S15 mRNA alternative secondary structure.

DISCUSSION

Secondary structure determination and representation are key steps of RNA analysis. Not only do these folding models serve as working hypotheses for their own structural refinements, but they should also lead to a better knowledge of higher structure interactions and to the prediction of structural features and molecular mechanisms involved in their function. ESSA is the first package which proposes an integration of several crucial aspects of RNA folding analysis (Fig. 1). A comprehensive visual representation of secondary structure constitutes the core part of ESSA around which are organized editing facilities, and a set of analysis functions allowing structural motif identification, thermodynamic calculation and secondary structure comparison. This makes ESSA the first program which integrates the three approaches currently used to predict secondary and tertiary contacts. Moreover, the communication protocol developed between ESSA and SAPSSARN enhances the predictive aspect of each of these two programs: ESSA benefits from the constraint satisfaction approach whereas SAPSSARN becomes more readable owing to a clear display of its results. Finally, a high level of interactivity, in a very natural and intuitive way, allows the user to drive each stage of a working session according to his knowledge.

The two methods implemented for displaying secondary structure cover the main aspects required by the biologists: RNA foldings are easy to produce, they highlight the main features of RNA and long molecules are easy to manage (Fig. 3). The complementary features of these two programs make them well adapted to different application fields. The fully automatic program (13) is suited to fast production of untangled secondary structure models. It is also helpful in removing overlaps with the interactive approach by giving a first readable untangled draft version. Nevertheless, it is sensitive to the complexity of the folding which can be defined as a function of (i) the number of multibranched loops, (ii) their size and (iii) the number of branches rooted on each of them. Thus it is possible that, given a configuration of the parameter set, no solution can be computed. This was the case for numerous sequences extracted from the datafile containing the universal core of secondary structure of the LSU rRNA. By contrast, the interactive approach (14) works whatever the length of the molecule and its folding complexity. It is better adapted to the evolutionary approach since it allows the representation of subdomains that are homologous among related molecules of different species with similar orientations in order to emphasize structural homologies even if long insertions/deletions interrupt conserved domains.

The palette of colors and symbols allows the user to create his own code to identify data as diverse as tertiary contacts, crosslinked nucleotides, inter-molecular interactions involving either RNA–protein or RNA–RNA and any other structural and functional features able to help in the interpretation of folding in terms of structure–function relationships. This integration of diverse knowledge increases largely the predictive value of the secondary structure model. For instance, the superposition of modified nucleotides along the universal core of secondary structure of rRNA molecules with the viewing of long complementarities with small nucleolar RNAs (snoRNAs) were at the basis of the recent finding of their guide function in the 2′-*O*-methylation of rRNA (27,28), a role which was later confirmed by *in vivo* experiments (28,29). These labelling functions are more particularly suited to visualizing probe accessibilities. Such information facilitates the determination of RNA secondary structure and the interpretation of computed solutions as illustrated in the case of S15 mRNA (Fig. 6). They become essential when chemical probing is not restricted to Watson–Crick pairings but is performed to identify the more diverse hydrogen bonds involved in tertiary interactions.

The interest of these editing and viewing tools for a better understanding of structure–function relationships was enhanced by the integration of several more advanced analysis functions. The searching sequence motif function takes advantage of the visual representation of labelled RNA folding to let the user estimate by eye, owing to his expertise, the significance of the occurrences. It allows the identification of complex structural motifs by searching sequentially each sequence segment of the query structural feature. The visual inspection of the structural environment of each occurence makes it relatively straightforward to estimate the drift with the searched one. In particular, insertion/deletions are easy to appreciate (Fig. 5) whereas they would have been difficult to take into account in an automatic approach such those developed to scan databases (30,31) or by a measure of similarity (32). Thus despite its simplicity, this method has a high predictive value, in particular through its coupling with aligned and structured datafiles, which increases the significance of occurences by integrating the constraints which are exerted during the course of evolution.

Using this approach we identified in the LSU rRNA a structural motif closely related to an 11 nt motif previously demonstrated to be involved in 3D interactions with a GAAA terminal loop in group I introns (22), group II introns (33) and in the RNA component of RNase P of *Bacillus megaterium* (34). Surprisingly, whereas this motif is present several times in group I introns, we found it only once along the entire LSU rRNA conserved core. This suggests a different mode of evolution for structural motifs involved in 3D interactions in rRNA, perhaps biased by the presence of numerous RNA binding proteins. The strong evolutionary constraint which is exerted on the key features of the LSU rRNA group I-like motif strongly supports its probable essential role in the elaboration of ribosomes or in their functioning. It could be a key feature in the 3D organization of this region which directly binds the ribosomal protein L23 (Fig. 5a) (35,36). Accordingly, the identification of its interacting partner would be an invaluable help. The differences observed in the receptor between the LSU rRNA and the group I intron could reflect variations in its interacting substrate. Alternatively, these differences could be induced by a contact with L23. A selex approach (37) associated with a search for covariating positions should help in identifying the substrate whereas NMR should reveal the precise spatial organization of this loop.

The most widely used tool for determining secondary structure folding is certainly the thermodynamic approach. Although it often fails to predict the overall folding of an RNA, this method brought important local information about the potential to form short stems or helix regions. ESSA, which gives the possibility of

modifying a folding, can also test alternative interactions in terms of free energy difference with the original model. Thus, by choosing among the various thermodynamic models proposed (38,39) and by adjusting, if necessary, the thermodynamic parameters, the user can compensate for the partial understanding of the different parameters acting on RNA folding. In contrast, the comparative approach relies on the construction of aligned sequence files followed by a search for compensated changes (40,41 for the most recent review). Once a secondary structure is determined for an aligned family of homologous sequences, it can be included in the alignment giving rise to a structured and specialized databank. Programs were recently developed to automatically increment new sequences in this particular format by aligning them using both sequence and secondary structure homologies (24). ESSA optimizes the use of these files through the production of four modes of secondary structure extraction and consensus display. The derivation of secondary structure models on the one hand and the identification of tertiary contacts on the other often necessitates the simultaneous use of information coming from the three approaches as complementary constraints to drive RNA folding. This has prompted us to develop communication between ESSA and SAPSSARN, to allow the user to participate directly in the computational folding of RNAs. He can thus manage interactively either within the ESSA display or within the SAPSSARN matrix any kind of structural constraint. These constraints are then propagated within the SAPSSARN matrix resulting in the elimination of forbidden pairs. For example, by removing pseudoknot constraints in SAPSSARN, these, if any, are visualized in ESSA as illustrated by S15 mRNA analysis (Fig. 6). Accordingly, this communication also allows us to address the question of the RNA 3D interactions and of the dynamics of the interactions by pointing to alternative interactions.

Our ultimate goal is to make ESSA a unique tool for analysing RNA; from their sequences to the production of 3D models. Toward this aim we are currently integrating programs devoted to the identification of tertiary contacts based on the high visualization potentialities of ESSA. We also plan to develop communications with programs dedicated to 3D RNA reconstruction. The selection function of ESSA will allow the extraction of basic structural features in order to build their 3D structures separately and then assemble these pieces of the 3D RNA puzzle (42).

ACKNOWLEDGEMENTS

We thank Dr Jean Pierre Bachellerie and Pr Eric Westhof for their constant interest and support. This work was financially supported in part by grants from the Groupement d'Intérêt Public, Groupement de Recherches et d'Etude sur les Génomes (GIP GREG) and from French Education and Research ministry to F.C. and B.M. (ACC SV 13 and 07).

REFERENCES

1 Jaeger,L., Michel,F. and Westhof,E. (1996) *Nucleic Acids Mol. Biol.*, **10**, 33–51.
2 Ehresmann,B., Ehresmann,C., Romby,P., Mougel,M., Baudin,F., Westhof,E. and Ebel,J.P. (1990) Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome, Structure, Function and Evolution.* American Society for Microbiology, Washington DC, pp. 148–159.
3 Kolchanov,N.A. Titov,I.I., Vlassova,E. and Vlassov,V.V. (1996) *Prog. Nucleic Acids Res. Mol. Biol.*, **53**, 133–191.
4 Zuker,M. (1989) *Science*, **244**, 48–52.
5 Woese,C.R. and Pace,N.R. (1993) in Gesteland,R.F. and Atkins,J.F. (eds) *The RNA World*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 91–117.
6 Gutell,R.R. (1993) *Curr. Biol.*, **3**, 313–322.
7 Westhof,E. and Michel,F. (1994) in Nagai,k. and Mattaj,I.W. (eds), *RNA–Protein Interactions*. Oxford University Press, pp. 25–51.
8 Major,F., Turcotte,M., Gautheret,D., Lapalme,G., Fillion,E. and Cedergren,R. (1991) *Science*, **253**, 1255–1260.
9 Gaspin,C. and Westhof,E. (1995) *J. Mol. Biol.*, **254**, 163–174.
10 Gouy,M. (1987) in Bishop,M.J. and Rawlings,C.J. (eds), *Nucleic Acid and Protein Sequence Analysis, A Practical Approach.* IRL Press, Oxford, Washington DC, pp. 259–284.
11 McCaskill,J. (1990) *Biopolymers*, **29**, 1105–1119.
12 Jaeger,J.A., Turner,D.H. and Zuker,M. (1990) *Methods Enzymol.*, **183**, 281–305.
13 Muller,G., Gaspin,C., Etienne,A. and Westhof,E. (1993) *Comput. Applic. Biosci.*, **9**, 551–561.
14 Perochon-Dorisse,J., Chetouani,F., Aurel,S., Iscolo,N. and Michot,B. (1995) *Comput. Applic. Biosci.*, **11**, 101–109.
15 Chiu,D.K.Y. and Kolodziejczac,T. (1991) *Comput. Applic. Biosci.*, **7**, 347–352.
16 Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
17 Gautheret,D., Damberger,S.H. and Gutell,R.R. (1995) *J. Mol. Biol.*, **248**, 27–43.
18 Maidak,B.L., Olsen,G.J., Larsen,N., Overbeek,R.O., McCaughey,M.J. and Woese,C.R. (1997) *Nucleic Acids Res.*, **25**, 109–110.
19 DeRijk,P., Van de Peer,Y. and De Wachter,R. (1997) *Nucleic Acids Res.*, **25**, 117–122
20 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
21 Matzura,O. and Wennborg,A. (1996) *Comput. Applic. Biosci.*, **12**, 247–249.
22 Costa,M. and Michel,F. (1995) *EMBO J.*, **14**, 1276–1285.
23 Philippe,C., Bénard,L., Portier,C., Westhof,E., Ereshmann,B. and Ehresmann,C. (1995) *Nucleic Acids Res.*, **23**, 18–28.
24 Corpet,F. and Michot,B. (1994) *Comput. Applic. Biosci.*, **10**, 389–399.
25 Hassouna,N., Michot,B. and Bachellerie,J.P. (1984) *Nucleic Acids Res.*, **12**, 3563–3583.
26 Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996) *Science*, **273**, 1678–1696.
27 Nicoloso,M., Qu,L.H., Michot,B. and Bachellerie,J.P. (1996) *J. Mol. Biol.*, **260**, 178–195.
28 Kiss-Laszlo,Z., Henry,H., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) *Cell*, **85**, 1077–1088.
29 Cavaillé,J., Nicoloso,M. and Bachellerie,J.P. (1996) *Nature*, **383**, 732–735.
30 Dandekar,T. and Hentze,M.W. (1995) *Trends Genet.*, **11**, 45–50.
31 Billoud,B., Kontic,M. and Viari,A. (1996) *Nucleic Acids Res.*, **24**, 1395–1403.
32 Chevalet,C. and Michot,B. (1992) *Comput. Applic. Biosci.*, **8**, 215–225.
33 Costa,M., Deme,E., Jacquier,A. and Michel,F. (1997) *J. Mol. Biol.*, **267**, 520–536.
34 Tanner,M.A. and Cech,T.R. (1995) *RNA*, **1**, 349–350.
35 Vester,B. and Garrett,R.A. (1984) *J. Mol. Biol.*, **179**, 431–452.
36 Egebjerg,J., Christiansen,J. and Garrett,R.A. (1991) *J. Mol. Biol.*, **222**, 251–264.
37 Costa,M. and Michel,F. (1997) *EMBO J.*, **16**, 3289–3302.
38 Turner,D.H., Sugimoto,N. and Freier,S.M. (1988) *Annu. Rev. Biophys. Chem.*, **17**, 167–192.
39 Antao,V.P. and Tinoco,I. (1992) *Nucleic Acids Res.*, **20**, 819–824.
40 Gutell,R.R. (1996) in Zimmermann,R.A. and Dahlberg,A.E. (eds), *Ribosomal RNA: Structure, Evolution, Processing and Function in Protein Biosynthesis.* CRC Press, Boca Raton, Florida, pp. 111–128.
41 Michel,F. and Costa,M. (1997) *RNA Structure and Function.* Cold Spring Harbour Laboratory Press, Cold Spring Harbor, NY, in press.
42 Westhof,E., Masquida,B. and Jaeger,L. (1996) *Folding Des.*, **1**, R78–R88.