ELSEVIER

Biophysical Chemistry 59 (1996) 179–184

Biophysical Chemistry

A graph-topological approach to recognition of pattern and similarity in RNA secondary structures

Giorgio Benedetti, Stefano Morosetti *

Dipartimento di Chimica, Università di Roma ''La Sapienza'', Piazzale A. Moro 5, 00185 Rome, Italy

Received 15 June 1995; accepted 15 September 1995

Abstract

Secondary and tertiary RNA structures play an important role in many biological processes. Therefore the necessity arises to find similar higher-order structures for different but functionally homologous RNA sequences. We propose here a graph-topological approach to the problem, which shows two main features: (a) A simplified graph representation which allows the recognition of similarity of RNA secondary structures with the same branching look despite minor differences. This allows comparison among foldings from different sequences, and ''pruning'' of the secondary structures not shared by all the sequences since the early stages of the search. (b) The graph representation is encoded by the Randić topological index, and the search for the folding similarity is reduced to checking the identity of single numbers. These characteristics make this approach significantly different, less depending on empirical criteria, and less computationally heavy then previous methods, where the folding consensus has been measured by an alignment procedure or correlation of strings representing the secondary structures. Some U2 snRNA and viroid sequences are studied by this approach, which is imbedded in our previous search method based on genetic algorithms.

Keywords: Chemical graph; Graph-topological approach; Randić topological index; RNA; Genetic algorithms

1. Introduction

1.1. Problem

RNA molecules have diverse biological functions that presumably require well-defined three-dimensional conformations [1]. The complexity of the life cycle (e.g. in the viroids) [2] or the implication in regulatory mechanisms [3], suggest in some cases the intervening of different tertiary structures. Predicting the secondary structure first, and then proceeding to the tertiary structure could be a fruitful approach.

The problem of determining RNA secondary structures is a multimodal problem, i.e. a great number of local solutions exists. Computing suboptimal foldings is necessary because of the uncertainties inherent in the thermodynamic data, and the loss of the three-dimensional interactions. Such research can be difficult because the number of the suboptimal foldings can be very large, and they are often local alternatives in the base pairing rather than structures very different in the branching.

The interpretations of the results are often ambiguous because of the number of RNA secondary

* Corresponding author.

SSDI 0301-4622(95)00119-0

structures with comparable free energies. Comparative studies with homologous molecules from different organisms add information. In fact, it is highly probable that the foldings essential for the function of the molecules are preserved during the evolution. Making use of this implicit information means to restrain the search to RNA secondary structures of general occurrence.

1.2. Approach

The main available computer approaches to the pattern similarity search use an alignment procedure [4] or the correlation [5] of strings representing the elementary structural "motifs", as they occur along the primary sequence. These approaches require a very large computational effort when they are used for searching for optimal and suboptimal similar foldings in many different sequences. In fact, the number of such comparisons increases with the square power of the number of sequences involved. Conversely it can be expected that the greater is the number of compared sequences, the bigger is the effectiveness in restraining the common foldings. Moreover, some difficulties can arise in the recognition of similarities when a very different number of bases, or of elementary structural motifs, occurs in the alignment or in the correlation procedure, respectively.

Searching for a new approach to overcome the above difficulties, it is first necessary to specify what is meant by the term "similar" when referring to the RNA foldings. The more "simple" definition is that structures are similar when they have the same branching look, i.e. the branching points have the same number of branches and are in the same topological relationship, despite having different lengths of branches. This way of looking at the problem suggests the representation of the secondary structures through structural graphs similar to the molecular ones, where the vertices are the loops, and the edges are the stems (see Fig. 1a and b) [6]. To overcome the problem of the length of the branches, the vertices connecting only two edges are omitted, so generating a "simplified" graph representation (see Fig. 1c).

The topological indices seem to be promising candidates for comparing the structural graphs. They are obtained by following mathematical rules applied to the graphs, with the aim of obtaining numbers characterised by properties such as: gradual change with gradual changes in the structures, good discrimination of isomers, and correct size dependence [6]. Many topological indices exist, and among them we choose the Randić's one, on the basis of the criteria more important for our problem: the correspondence between the graph and the index value being the more possibly univocal, and the simplicity of calculus.

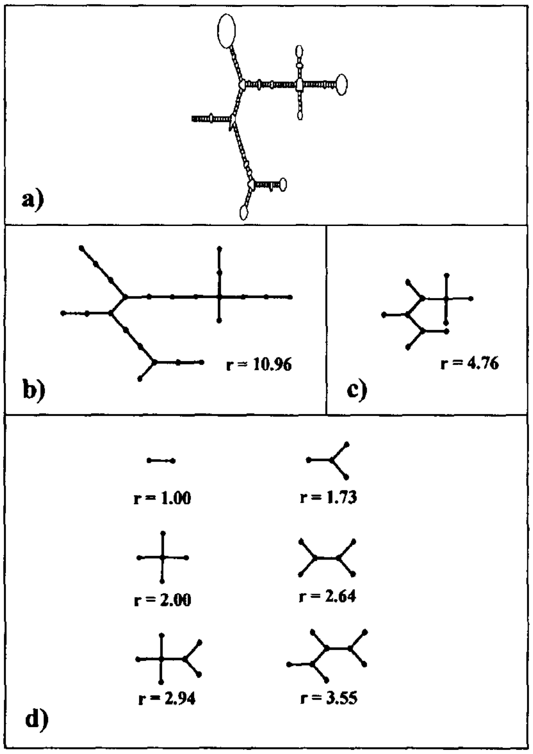

Fig. 1. RNA secondary structure, its graphical representation, and its Randić topological index. (a) Putative suboptimal free energy secondary structure. (b) Structural graph representation: the points (vertices) represent the loops, and the edges represent the helices. (c) Simplified structural graph representation: the vertices connecting only two edges are omitted. (d) Some of the graphical representations in increasing order of complexity. The Randić topological indices (r) are reported.

The Randić index is defined as:

$$\chi = \sum_{\text{all edges}} [d(i)d(j)]^{-0.5} \qquad (1)$$

where the valency of a vertex is the number of edges incident to it, and $d(i)$ and $d(j)$ are the valencies of the vertices i and j that define the edge ij. The Randić indices of some structural graphs are shown in Fig. 1b, c and d.

The comparison of two secondary structures is obtained by the comparison of their simplified structural graphs, and this is accomplished by simply checking the identity of their Randić indices.

2. Method

The above-stated graph-topological approach allows to encode the structural information from the Randić topological index. Therefore the search for similar foldings is reduced to the request of the same index value for structures arising from different sequences. This criterion can be easily implemented in any algorithm to search for the similarity consensus.

The filtering action will be accomplished by accepting the foldings of different sequences that exhibit the same Randić indices, and by discarding the other ones. This action can be performed anywhere where the secondary structures are defined in the algorithm. In the recursive algorithms the secondary structures are built in the final stage, and just then the topological indices can be calculated. Conversely, the combinatorial ''tree'' searches and genetic algorithms manipulate the secondary structures in all the stages of the search, and therefore they can extensively use the filtering action of the graph-topological approach, by discarding the tentative or intermediate solutions, which do not exhibit the same Randić index. In these cases there is the considerable advantage of addressing the search toward the common foldings, by using the information that the same performed biological function probably involves similar foldings (see Section 1.1), and reducing the computational effort. The use of the simplified graph representation is recommended in such cases, since it allows the comparison of foldings for structures at a different degree of completeness, because of its independence from the length of the branches.

2.1. Genetic algorithms and graph-topological approach

We embedded the above-stated graph-topological approach in our previously developed search method based on the genetic algorithms. We recall here only the main ideas of the genetic algorithms in general, and of our application to the RNA folding in particular, referring to the pertinent article [7] for a complete description.

Genetic algorithms are a search method used for solving problems by selection, recombination, and mutation of tentative solutions, until the better ones are achieved. This strategy mimics the fundamental rules of natural genetic evolution. The main steps of our algorithm to search for the RNA foldings include the following:

- Starting population: the foldings are generated by a random choice of compatible helixes.
- Fitness function: each folding is estimated principally by its free energy, calculated by following Freier's rules [8].
- Reproduction: it is the random choice of the foldings, weighted by the fitness function.
- Cross-over: it is the generation of two new foldings obtained by the exchange of ''pieces'' of structure from two old foldings. It is applied over a fraction of the reproduced individuals.
- Mutation: it is the random change in the helixes forming a folding.
- Iteration: the individuals (foldings) obtained through the above operators, constitute a new population that substitutes the old one, after which the process is repeated.

The next subsections deal only with the main modifications brought on our search method to achieve a parallel search on the sequences, and to perform the topological selection of the foldings.

2.1.1. Parallel search

The search is arranged stepwise: at each step one iteration of the genetic algorithm is performed on all the sequences under consideration. At the end of the step the Randić indices are calculated, and the ''pruning'' of the foldings is performed (see Section 2.1.3).

2.1.2. Mutation

During each step the mutations of the duplicate structures are performed with the requirement of preserving the Randić index. This request is very effective in avoiding an excessive increase in the number of different Randić indices, often not common for all the sequences. Therefore it consequently shortens the time required for the search.

2.1.3. Pruning the foldings

At the end of each step the Randić indices common for all the sequences are determined. The foldings that present different Randić indices are deleted, and they are replaced by new structures randomly generated, as happened at the beginning of the search.

2.1.4. Iteration

The reproduction, cross-over, mutation, injection, and pruning steps are iterated until stable, best and mean values of the fitness function based on free energy values [7] are obtained for each "survived" Randić index. We used the free energy calculated by following Freier's rules [8].

2.2. Drawings

We used the computer program SQUIGGLES of the GCG software package [9,10] to draw some parts of our figures.

3. Results and discussion

We have performed two parallel searches of the optimal and suboptimal common foldings by using (a) four U2 snRNA sequences and (b) two RNA viroid sequences.

We used the simplified graph representation of the foldings to recognise the similarity, thereby overcoming the problems of the different length of the sequences, and of minor differences in the positions of the loops along the branches. The graph representation is encoded as Randić index, and the comparison is reduced to checking the identity of Randić indices.

3.1. U2 snRNA sequences

The U2 snRNA sequences chosen have substantial differences in their base sequences and lengths. They are: Human U2 snRNA sequence [11]; *Caenorhabditis Elegans* U2 snRNA sequence, which was chosen because of its difference from Human U2 in the base sequence [12]; *Saccharomyces cere-*

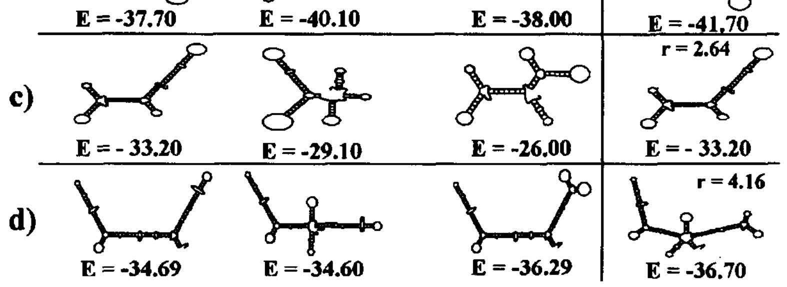

Fig. 2. U2 snRNA sequences. (a) Human. (b) *Caenorhabditis Elegans*. (c) *Trypanosoma congolense*. (d) *Saccharomyces cerevisiae*. The optimal-free-energy secondary structures are shown in the right column. Randić topological indices (r) of the reduced graph representation are reported. The common foldings are shown in the three columns on the left. The common Randić topological indices (r) of the reduced graph representation are reported at the top. Free energies (E) are given in kcal/mole.

visiae U2 snRNA sequence, of which we studied the fragment that contains the biological activity [13], which is still longer than human U2; *Trypanosoma congolense* U2 snRNA sequence [14], which is considerably shorter than human U2.

A population of 100 foldings for each sequence and 100 iterations were sufficient to reach convergence in the results. Only three common foldings were found, and they are shown in Fig. 2.

If only the common foldings are examined, the followings facts could be of some significance: (a) the minimum free energy is at the same Randić value of 3.55 for the two sequences which process nuclear mRNAs by *cis*-splicing (that is Human and *Saccharomyces cerevisiae* U2 snRNA sequences); (b) the minimum free energy is at the Randić value of 2.64 for the *Trypanosoma congolense* U2 snRNA sequence, which processes nuclear mRNAs by *trans*-splicing; (c) the foldings with Randić values of 3.55 and 2.64 have practically the same free energies in the *Caenorhabditis Elegans* U2 snRNA sequence, for which both *cis*- and *trans*-splicing are known to occur.

3.2. Viroid RNA sequences

The viroids are single-stranded RNA molecules [15] that can go through different suboptimal foldings during the various steps of their complex life cycle. We choose two viroids with substantial differences in the lengths and base sequences. They are apple scar skin viroid (ASSV) sequence [16] and *Coleus blumei* viroid (CBVD) sequence [17].

A population of 1000 foldings for each sequence and 200 iterations were sufficient to reach convergence for the results. Only three common foldings were found and they are shown in Fig. 3.

3.3. Concluding remarks

The main ideas used in the parallel search for homologous foldings are the graph representation of the foldings, and their identification through the Randić topological index. This procedure allows to encode the structural information in a number, and this makes the comparison among the foldings very simple. When the reduced graph representation is used, the comparison allows the recognition of the similarity among structures with the same branching look, despite their having minor differences in the length or in the number of helixes along the single branch.

This approach can be easily implemented in any search method, but it is very effective when pruning of the structures during the search is possible, which is therefore addressed towards the common foldings.

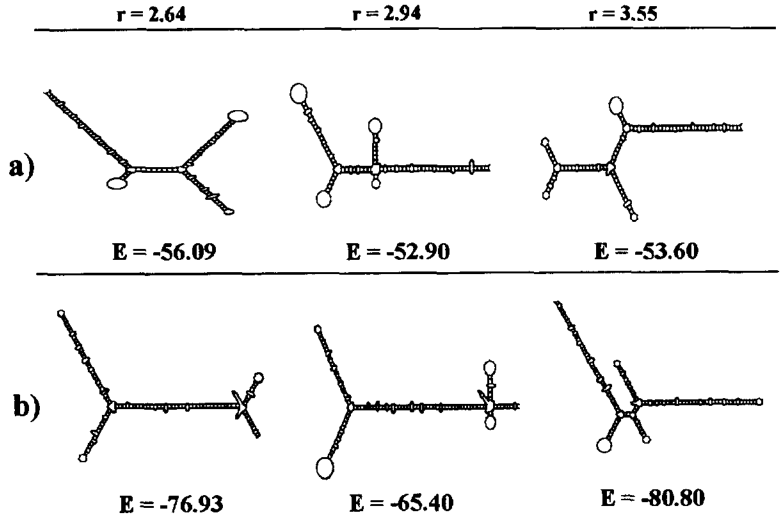

Fig. 3. Viroid sequences. (a) Apple scar skin viroid. (b) *Coleus blumei* viroid. The common foldings are shown. The common Randić topological indices (r) of the reduced graph representation are reported at the top. Free energies (E) are given in kcal/mole.

Our search method, based on genetic algorithms, seems to be particularly suitable to accommodate the graph-topological representation, and the successful parallel RNA searches shown here support this opinion.

Acknowledgements

We thank Professor P. De Santis, Dr. A. Scipioni, and Dr. R. Mastrangeli for helpful discussions. This work was supported by Ministero della Università e della Ricerca Scientifica e Tecnologica, Grant 92 20902 444.

References

[1] B. Wimberly, G. Varani and I. Tinoco, Jr., Curr. Opinions Struct. Biol., 1 (1991) 405–409.

[2] T.O. Diener, Proc. Natl. Acad. Sci. USA, 83 (1985) 58–62.

[3] P.M. Bingham, T.-B. Chou, I. Mims and Z. Zachar, Trends Genet., 4 (1988) 134–138.

[4] D.A.M. Konings and P. Hogeweg, J. Mol. Biol., 207 (1989) 597–614.

[5] G. Benedetti and S. Morosetti, Eur. J. Biochem., 202 (1991) 241–248.

[6] Z. Mihalić and N. Trinajstić, J. Chem. Educ., 69 (1992) 701–712.

[7] G. Benedetti and S. Morosetti, Biophys. Chem., 55 (1995) 253–259.

[8] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Nielson and D.H. Turner, Proc. Natl. Acad. Sci. USA, 83 (1986) 9373–9377.

[9] J. Devereux, P. Haeberly and O. Smithies, Nucleic Acids Res., 12 (1984) 387–395.

[10] P. Hogeweg and B. Hesper, Nucleic Acids Res., 12 (1984) 67–74.

[11] E.O. Shuster and C. Gutrie, Nature, 345 (1990) 270–273.

[12] J. Thomas, K. Lea, E. Zucker-Aprison and T. Blumenthal, Nucleic Acids Res., 18 (1990) 2633–2642.

[13] A.H. Igel and M. Ares, Jr., Nature, 334 (1988) 450–453.

[14] C. Tschudi, S.P. Williams and E. Ullu, Gene, 91 (1990) 71–77.

[15] P. Keese and R.H. Symons, in T.O. Diener (Ed.), Physical-Chemical Properties: Molecular Structure (Primary and Secondary), The Viroids, Plenum Press, New York, 1987, pp. 37–62.

[16] H. Puchta, R. Luckinger, X. Yang, A. Hadidi and H.L. Saenger, Plant Mol. Biol., 14 (1990) 1065–1067.

[17] R.L. Spieker, Y.C. Haas Charng, K. Freimuller and H.L. Saenger, Nucleic Acids Res., 18 (1990) 3998–3998.