BioSystems, 27 (1992) 85–96
Elsevier Scientific Publishers Ireland Ltd.

Neural computing in discovering RNA interactions

Yoshiyasu Takefuji[a], Dora Ben-Alon[b] and Arieh Zaritsky[b]

[a]*Department of Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, Ohio 44106 (USA) and* [b]*Department of Life Sciences, Ben-Gurion University of the Negev, P.O. Box 653, Be'er-Sheva 84105 (Israel)*

(Received January 15th, 1992)
(Revision received May 16th, 1992)

High-order RNA structures are involved in regulating many biological processes; various algorithms have been designed to predict them. Experimental methods to probe such structures and to decipher the results are tedious. Artificial intelligence and the neural network approach can support the process of discovering RNA structures. Secondary structures of RNA molecules are probed by autoradiographing gels, separating end-labeled fragments generated by base-specific RNases. This process is performed in both conditions, denaturing (for sequencing purposes) and native. The resultant autoradiograms are scanned using line-detection techniques to identify the fragments by comparing the lines with those obtained by 'alkaline ladders'. The identified paired bases are treated by either one of two methods to find the foldings which are consistent with the RNases' 'cutting' rules. One exploits the maximum independent set algorithm; the other, the planarization algorithm. They require, respectively, n and n^2 processing elements, where n is the number of base pairs. The state of the system usually converges to the near-optimum solution within about 500 iteration steps, where each processing element implements the McCulloch-Pitts binary neuron. Our simulator, based on the proposed algorithm, discovered a new structure in a sequence of 38 bases, which is more stable than that formerly proposed.

Keywords: High-order RNA structures; Biological regulation; Artificial intelligence; Neural nets; Maximum independent set; planarization algorithms.

1. Introduction

The four common bases of an RNA molecule are cytosine (C), uracil (U), adenine (A) and guanine (G). The double-helix of an RNA forms when two separate sections (with a 5′-end to 3′-end polarity) become linked together in an anti-parallel manner by weak hydrogen bonds between specific, complementary bases: A always pairs with U and G pairs with C. The primary structure of RNA is defined as its linear base sequence. The secondary structure is determined by its folding into a two-dimensional shape. Folding into a three-dimensional shape is called tertiary structure and structures formed by interactions with other molecules are quaternary.

RNA molecules are involved in a wide range of functions in the living world, exerted in part by the three-dimensional conformations to which they can fold (e.g., Zaritsky et al., 1988; Dahlberg and Abelson, 1989 and 1990; Puglisi et al., 1991). The stability of a structure is measured by the free-energy difference between folded and unfolded forms. An RNA sequence can often form alternate structures of similar stabilities, which may be the reason for its role in various processes (e.g., translation). Predictions of the secondary structure of RNA, that is, its base-pairing pattern (whether based on free energy calculations or distance geometries or inferred from compensatory mutations; e.g., Zuker and Stiegler, 1981; Williams and Tinoco, 1986; Martinez, 1988; Zuker, 1989a,b; Mei et al.,

Correspondence to: Yoshiyasu Takefuji, Department of Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, Ohio 44106, USA.

0303-2647/92/$05.00 © 1992 Elsevier Scientific Publishers Ireland Ltd.
Printed and Published in Ireland

1989; Takefuji et al., 1990a and b; Major et al., 1991) are more reliable than those predicted for proteins from amino acid sequences (Karplus and Petsko, 1990).

1.1. Current approaches to predicting secondary RNA structures

Fresco (Fresco et al., 1960) used the first model to predict secondary structures in RNA. Two types of algorithms have been reported: the combinatorial method (Pipas and McMahon, 1975) and the recursive (or dynamic programming) method (Nussinov et al., 1978). Both algorithms, as well as the latest method proposed by Zuker (1989b), are based on sequential computation. Unfortunately, few parallel algorithms based on molecular thermodynamics models have been reported. Recently, Qian and Sejnowski (1988) and Holley and Karplus (1989) have reported a backpropagation algorithm using a three-layer-feed-forward neural network for protein secondary structure prediction. Their method is based on the correlation between secondary structure and amino acid sequences, but has the following drawbacks as compared with the conventional RNA folding algorithms based on molecular thermodynamics models: (1) they need a teacher to force the network to learn the correlation between a secondary structure and an amino acid sequence; (2) they cannot provide an accurate prediction if a completely uncorrelated new datum is given where the previously learned correlation information is useless; (3) their feed-forward neural network requires a prohibitively long learning process to deal with a long sequence of bases for RNA secondary structure prediction; (4) no theorem is given to determine the neural network architecture including how many hidden layers and how many hidden neurons per hidden layer should be used. Our algorithms (Takefuji et al. (1990a and b); and see Section 2) requires neither a teacher nor a learning process. The proposed maximum independent set parallel algorithm can yield the suboptimum solution within several hundred iteration steps using n processors (where n is the number of possible base pairs).

Generally speaking, the existing algorithms can be classified into three quite different approaches. One approach is a phylogenetic structure analysis of homologous RNAs that depends on multiple alignment of the molecules (Jaeger et al., 1989; Le and Zuker, 1991). Helix conservation is scored by a ratio of the number of times that the helix occurs in suboptimal foldings. The combination of compatible helices generates a secondary structure of the statistically more significant ones.

The property of RNA molecules to fold has yielded algorithms of the second approach, that compute optimal foldings with mathematical tools, based on either maximizing the number of pairings or minimizing the free-energy. Recursive algorithms have been used and a collection of secondary structures that can be found close to the energy minimum are generated (Le and Zuker, 1991). Using Turner et al.'s (1987) energy rules, the computer prediction accuracy was elevated to 70%. Predictions made within 10% of the lowest free energy include in them up to 90% of the phylogenetically known helices. Of course, the main problem remains to be solved, i.e., choosing the correct structure occurring in any given natural RNA. In addition, algorithms of this type require computer time proportional to n^m, where n is the number of variables and m the number of permitted values.

A third, new approach, that of neural representation, encodes the problem using artificial neurons (Steeg, 1989; Takefuji et al., 1990a and b). A fired neuron represents a possible base pair like G-C and A-U. This method resides in the energy family and the idea is to find the largest number of base pairs which will prove to have the minimum energy. Use of graph theory solution for finding the largest planar subgraph or the maximum independent set permit finding the largest set base pairs. Among the diverse structures, the more knowledge and rules from experts embedded in the neural network, the better is the solution and the faster it is obtained.

These models, albeit theoretical, do consider actual information available about chemical interactions between bases. For example, the nearest neighbor model approximates the stability of an RNA duplex by the sum of the free energy increments for all its ten nearest neighbors in the duplex provided by various measurements (e.g., Freier et al., 1986; Turner et al., 1987; Jaeger et al., 1989). However, the algorithms based on free energy minimization or distance geometry are limited because solutions might represent local minima, depending on the input structure, rather than the global minimum. There can therefore be numerous foldings within 5% to 10% of the computed minimum free energy.

Uncertainties and difficulties of these kinds can be mitigated by incorporation of additional data. For instance (Zuker, 1989b), incorporating nuclease data which identify single or double stranded regions results in a dot plot with compatible base pairs only. Further information is gained by determination of a common RNA secondary structure within a set of homologous RNAs (Jaeger et al., 1989; Le and Zuker, 1991) and analysis of time intervals in structural reconstructions (because both the building of an mRNA molecule and its passage into the cytoplasm of eukaryotic cells start from its 5′ end; Gultyaev, 1991). All of these procedures are not sufficient to yield the correct, bio-active structure occurring for any given RNA; the main problem is thus not yet resolved. Some interactions between unpaired bases of a folded RNA and between RNA or other regulatory molecules in a living cell complicate the achievement of meaningful conclusions still further (Garrett et al., 1981; Goringer and Wagner, 1988).

2. Maximum independent set (MIS) programs

2.1. The algorithm

An independent set in a graph is a set of vertices, no two of which are adjacent. A maximum independent set (MIS) is an independent set whose cardinality is the largest among all independent sets. The problem of finding an MIS for arbitrary graphs is non-deterministic polynomial-complete. The MIS problem, the max cut problem and the maximum clique problem are all interrelated with each other for finding ground states of spin glasses with exterior magnetic fields and solving circuit layout design problems in VLSI circuits and printed circuit boards.

Our parallel algorithm (Takefuji et al., 1990b) generates a near-MIS of a circle graph in nearly constant time. For an n-edge problem, the algorithm uses n processing elements, each implementing the McCulloch-Pitts (1943) neuron model.

Consider the simple circle graph (Fig. 1a) with 14 vertices and 7 edges. Figure 1b shows its adjacency graph, generated by edge-intersection in the circle graph. For example, the edge 'd' intersects with three edges: c, e and f. The adjacency graph $G(V, E)$ of the circle graph is given by $V = \{a, b, c, d, e, f, g\}$ and $E = \{(a\ b), (b\ c), (b\ e), (b\ f), (c\ d), (d\ e), (d\ f), (e\ f), (f\ g)\}$. A set of vertices $\{a, c, e, g\}$ in Fig. 1c is the MIS of this circle graph which is equivalent to the maximal planar subgraph, as shown in Fig. 1d. In other words, finding the MIS in a circle graph is equivalent to finding its maximum planar subgraph. In order to find the near-maximal planar subgraph in the circle graph with m-vertex and n-edge, n neurons (processing elements) are used in our algorithm. The output state of the ith neuron $V_i = 1$ means that the ith edge is not embedded in the circle graph. The state of $V_i = 0$ indicates that the ith edge is embedded in the circle graph.

The motion equation of the ith neuron for $i = 1,\ldots,n$ is given by:

$$\frac{dU_i}{dt} = A\left(\frac{\sum_{j=1}^{n} d_{ij}(1 - V_j)}{\text{distance }(i)}\right)(1 - V_i) - Bh\left(\sum_{j=1}^{n} d_{ij}(1 - V_j)\right)V_i \qquad (1)$$

where $d_{xy} = 1$ if the xth edge and the yth edge intersect each other in the circle graph, 0 otherwise. Note that A and B are constant

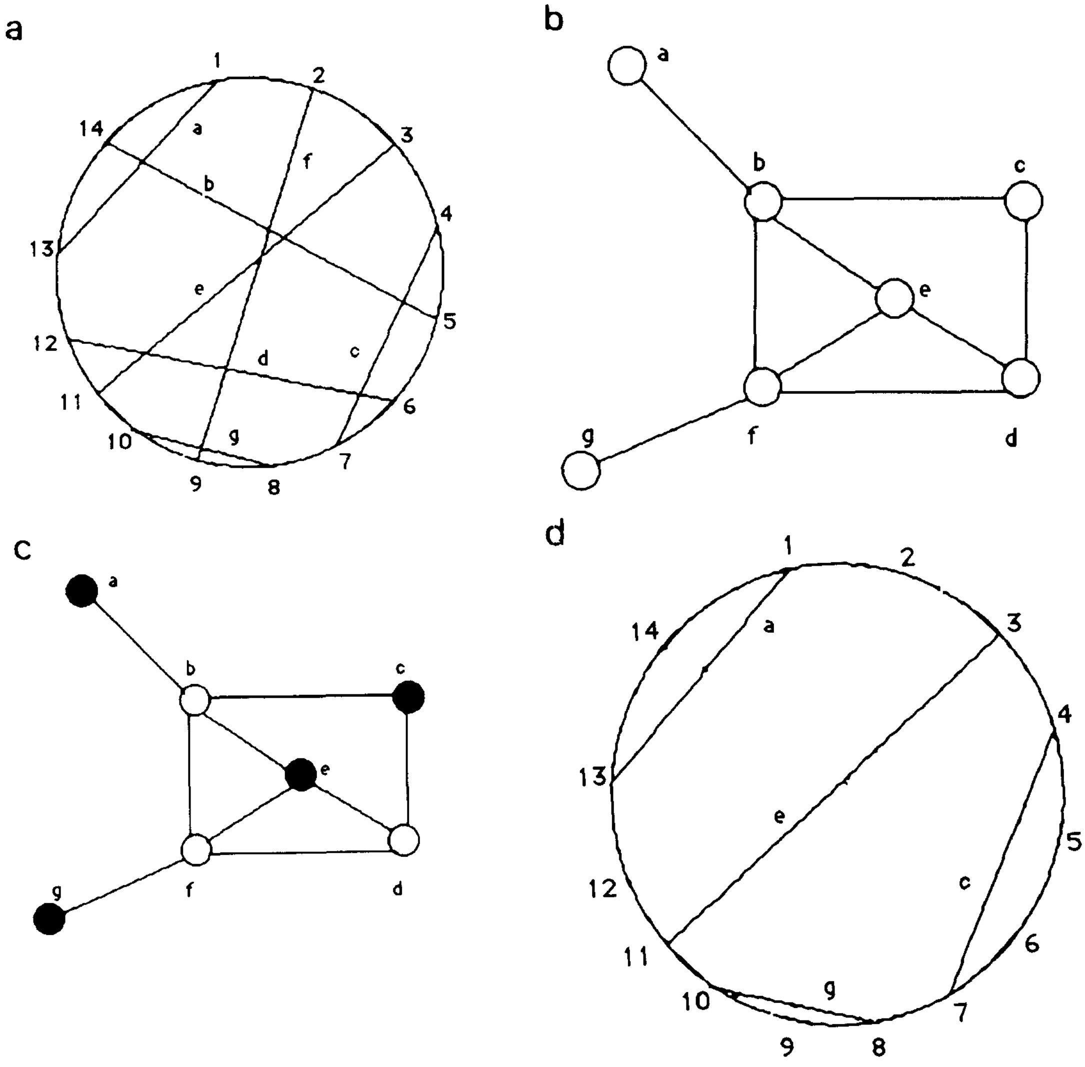

Fig. 1. (a) Circle graph with 14 vertices and 7 edges.
(b) Adjacency graph of Fig. 1a.
(c) Maximum independent set of Fig. 1a.
(d) Maximum planar subgraph of Fig. 1a.

coefficients. Edge-intersection conditions between the ith and the jth edges in the circle graph are given by: head $(i) <$ head $(j) <$ tail $(i) <$ tail (j) and head $(j) <$ head $(i) <$ tail $(j) <$ tail (i) where tail (i) and head (i) are two end vertices of the ith edge. Note that distance (i) is given by distance $(i) = \min(|\mathit{head}(i) - \text{tail}(i)|, |n + \text{head}(i) - \text{tail}(i)|)$ where tail $(i) >$ head (i) is always satisfied.

The function $h(x)$ is 1 if $x = 0$, 0 otherwise.

The first term is the inhibitory force in order to remove the edges which intersect with the ith edge in the circle graph. If the ith edge is removed from the circle graph, the first term will not be activated at all, because the state of the ith neuron $V_i = 1$. In order to keep the ith edge in the circle graph, the first term should not have any edge-intersection violation. Whenever the *i*th edge has any edge-intersection violation, it will tend to be eventually removed from the circle graph. The last term is the encouragement force to embed the *i*th edge in the circle graph. If the *i*th edge is removed but does not intersect with any other edges, the last term will force the *i*th neuron to be $V_i = 0$. In other words, the *i*th edge is encouraged to exist in the circle graph.

Our goal is to maximize the number of edges in the planar circle graph where an edge represents a possible base-pair (G-C or A-U). The goal of maximizing the number of base pairs for predicting the secondary structure in RNA viroids was supported by Diener (1987).

The motion equation of Eqn. 1 is slightly modified for predicting secondary structures of RNAs: (a) edge-intersection violation conditions must be updated. Six conditions to describe the edge-intersection in the circle graph are required:

head (i) < head (j) < tail (i) < tail (j), head (j) < head (i) < tail (j) < tail (i), tail (i) = tail (j), tail (i) = head (j), head (i) = head (j) and head (i) = tail (j).

The last four violation conditions are newly added to the first two, because a single base cannot be involved in more than one base-pair. The other modification is in the distance (i) function of Eqn. 1, where it is given by distance (i) = |head (i) − tail (i)|.

A sequence of m bases is given to the simulator. It generates the circle graph with m vertices and n edges where n is the number of possible base pairs. Each base-pair must also satisfy the hairpin-loop constraint, |head (i) − tail (i)| > 3, because it is sterically impossible to organize the hairpin loop with less than three bases. The circle graph is fed to the neural network simulator in order to find the near-MIS.

2.2. An example

The stability number for a given RNA secondary structure is the sum of the contributions of the loops, bulges and helices. The structure with the highest number is the most stable, called optimal folding. A sequence of 38 bases from residues 1118−1155 of *Escherichia coli* 16S rRNA served as an example to validate our simulator. Figure 2a shows the secondary structure proposed by Stern (Stern et al., 1988), where the structure stability (+7) is computed based on Tinoco's values:

(I). A-U pair, +1;
(II). G-C pair, +2;
(III). G-U pair, 0;
(IV). hairpin loops, -5 to -7;
(V). interior loops, -4 to -7;
(VI). bulges, -2 to -6.

Figure 2b shows the circle graph with 38 vertices and 151 edges, each edge represents a possible base-pairing.

When $A = B = 1$ and $U_i(0) = -5$ for $i = 1,...151$, the state of the system converged to the solution containing 14 edges (Fig. 2c) in the 104th iteration step. The secondary structure of the simulation result is given in Fig. 2d. Its stability number is +11, demonstrating that the simulator can find more stable structure than found by other means (Stern et al., 1988).

3. Planarization and RNA structure prediction

3.1. The algorithm

The mathematical problem to compute an optimal folding based on free-energy minimization is mapped onto a graph planarization problem (Takefuji et al., 1990b). We want to maximize the number of edges in a plane with no two edges crossing each other. The A-U or G-C base pairs are only considered as possible edges to be embedded in a plane while the bases are the vertices. In other words, for a fragment stretching from ribonucleotides i to j, it is denoted by the subscript ijth neuron where the output and the input is depicted by V_{ij} and U_{ij}, respectively for $i = 1,...,n-1$ and $j = i+1,...,n$.

Consider a sequence of fifteen bases (Fig. 3a). In our algorithm a single-row representation is used where five edges are embedded. A

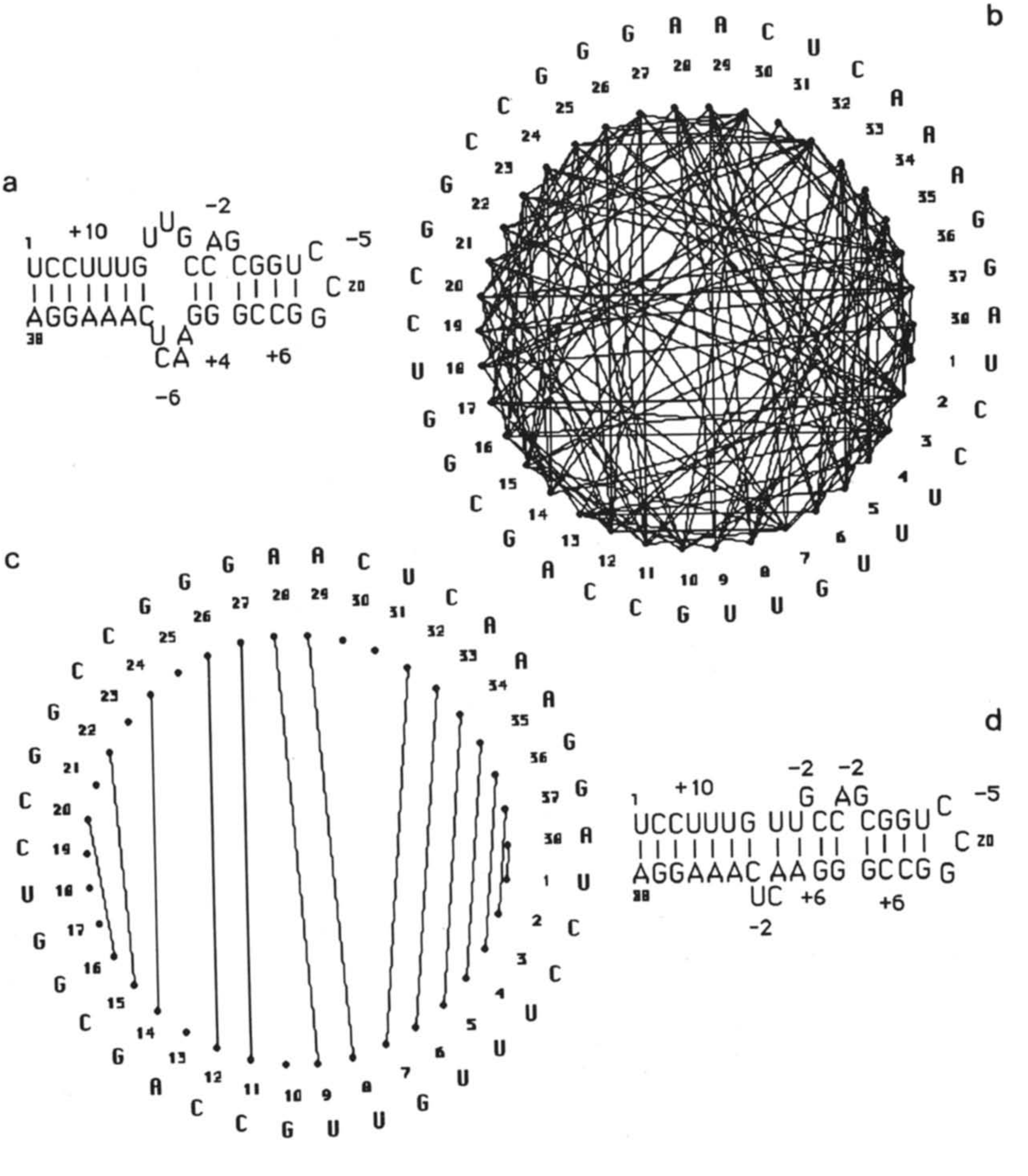

Fig. 2. (a) Secondary structure of a sequence of 38 bases, proposed by Stern (Stern et al., 1988).
(b) Circle graph with 38 vertices and 151 edges.
(c) State of the system after 104 iterations.
(d) Secondary structure predicted by our algorithm (Takefuji et al., 1990b).

(15 × 14)/2 neural network array (Fig. 3b) is used to predict the secondary structure of this problem. The following five functions are considered (Eqn. 2): $b(i,j)$, $g(i,j,k)$, $f(i,j)$, $p(i,j,t)$ and $h(x)$. The function $b(i,j)$ denotes the possible pairing: $b(i,j) = 1$ if i and j are one of the four legitimate base pairs (G-C, C-G, A-U, or U-A), 0 otherwise. Cross bonding is sterically impossible so that the graph must be planar without two edges crossing each other. A violation function $g(i,j,k) = 1$ if $i < j < k$, 0 otherwise has been described (Takefuji and Lee, 1989). The function $f(k,l)$ indicates the strength of a base pair bond between k and l bases: $f(k,l) = 2$ if k and l bases are a G-C pair, 1 if they are an A-U pair, 0 otherwise (Tinoco et al., 1971). The hairpin loop constraint is also considered in our algorithm where more than two bases are required to make a hairpin loop: the function $p(i,j,t) = 1$ if (j-hairpin) $< i$, 0 otherwise where hairpin is given by hairpin $= 4$ if $t = 0$, 55-t if 55-$t > 4$, 4 otherwise. The hill-climbing function is $h(x)$, 1 if $x = 0$, 0 otherwise.

To predict suboptimal foldings of a sequence of n bases, $n(n - 1)/2$ neurons are required. The motion equation of the ijth neuron is given by:

$$\frac{dU_{ij}}{dt} = -A_1\left(\sum_{k \neq i} V_{ik} - 1\right)b(i,j)$$
$$- A_2\left(\sum_{k \neq j} V_{jk} - 1\right)b(i,j)$$

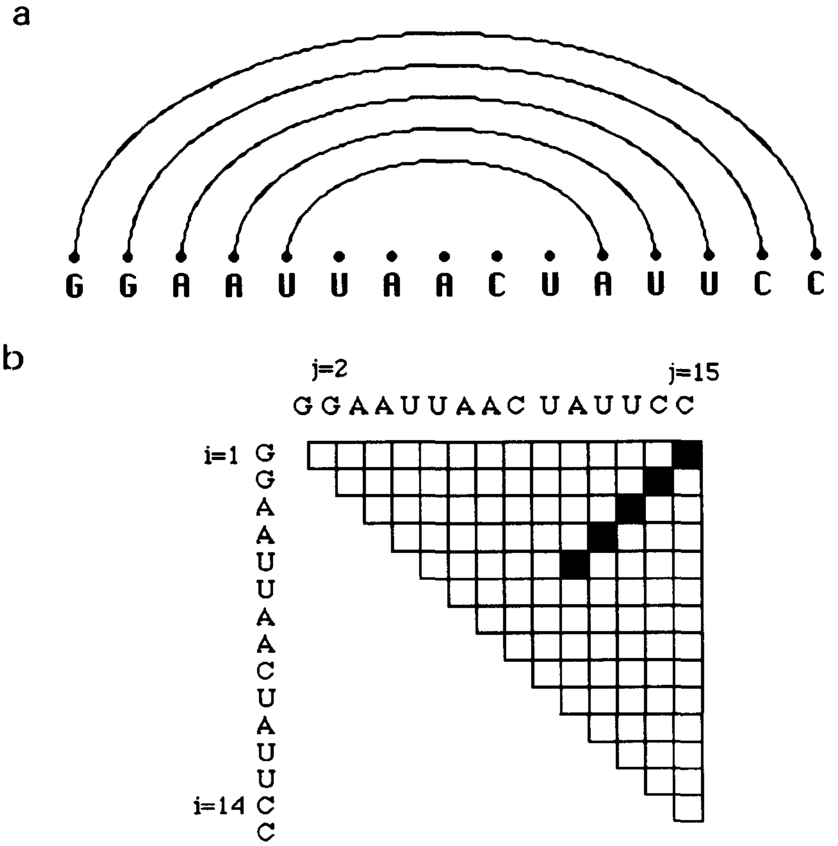

Fig. 3. (a) Single-row representation of 15 bases.
(b) Neural representation of a sequence of 15 bases (Fig. 3a).

$$
\begin{aligned}
&- B_1 \sum_{k < i < l < j} V_{klg}(k,i,l)g(i,l,j)f(k,l) \\
&- B_2 \sum_{i < k < j < l} V_{klg}(i,k,j)g(k,j,l)f(k,l) \\
&- Cp(i,j,t) \\
&+ Dh\left(\sum_{k \neq i} V_{ik}\right) \qquad (2)
\end{aligned}
$$

The first two terms force the *i*th and *j*th bases, respectively, to have one and only one bond. Note that if the *i*th base has strong violations caused by other bases, it cannot have any bond. The third and fourth terms are inhibitory and always satisfy planarization conditions. The fifth term is the inhibitory hairpin constraint which prohibits less than three bases to make a hairpin loop. The hill-climbing force term ($h(x)$) allows the state of the system to escape from the local minimum.

3.2. An example

Consider a sequence of 55 bases from R17 viral RNA (Tinoco et al., 1971). 1485 (= 55 × 54/2) neurons are used to solve this problem. Figure 4a and b, shows the state of the system after the 1st and the 61st iterations, respectively. The latter is translated to the predicted secondary structure in Fig. 4c. The total stability of the structure is +7, which is equivalent to free energy ΔG = -8.4 kcal per mol. When pairs (A^{15} to U^{41}), (A^{16} to U^{40}) and (A^{17} to U^{39}) are shifted to (A^{15} to U^{40}), (A^{16} to U^{39}) and (A^{17} to U^{38}), respectively, the total stability of the structure becomes +8, which is the optimum (Tinoco et al., 1971).

4. Real life analyses

4.1. Experimental (Wet) procedures

There exist several physico-chemical experimental procedures (e.g., Noller, 1984) to derive an RNA secondary structure, or to choose among alternative possibilities obtained by the conventional computer methods (Section 1.1). X-ray diffraction analysis (Holbrook et al., 1978), electron-microscopy (Jacobson et al., 1985) and absorption spectrum analysis (Reid, 1981), are some examples of direct methods, each with its own disadvantages. In all the indirect methods, the RNA molecule to be studied must be exposed to some chemical or enzymatic treatment and the nature and composition of products are analysed. These include: use of psoralens to cross-link the two DNA strands (e.g., Cimio et al., 1985), binding of synthetic oligonucleotides (e.g., Mankin et al., 1981) and the use of specific chemicals or specific ribonucleases (e.g., Ehresmann et al., 1987).

The only way to tell which of the candidate structures really exists is by probing it chemically. Strong evidence supporting the existence of a particular structure is usually obtained by analyzing products generated by ribonucleases with known structure-specific activities (Knapp, 1989). The molecule under scrutiny is labelled at either its 5′ or 3′ terminal nucleotide with ^{32}P and then submitted to partial digestion by each of a battery of specific RNases (for the minimum number of enzymes, see Zaritsky and Forester, 1991), under conditions which are assumed not to interfere with the native RNA conformation. Lengths of the labelled fragments thus generated are determined by gel electrophoresis and autoradiography, with a ladder containing all alkaline hydrolysis products of the RNA as size markers. These lengths identify the positions of enzymatic cleavages and hence the local secondary structure (Pieler et al., 1986; Gerhart et al., 1986; Ehresmann et al., 1987). (If the actual sequence of the given RNA molecule is unknown, the same procedure can be exploited to determine it, but with base-specific single-stranded RNases operating in denaturing conditions.)

Deciphering the results of such experiments are however tedious and not always fruitful. Algorithms that would allow computers do this job are desirable but unavailable. Here, we attempt to present one (Section 3), which is based on the third, neural network approach (1.1)

4.2. The 'reverse' (simulation) approach

A parameterizable simulation program (GEL)

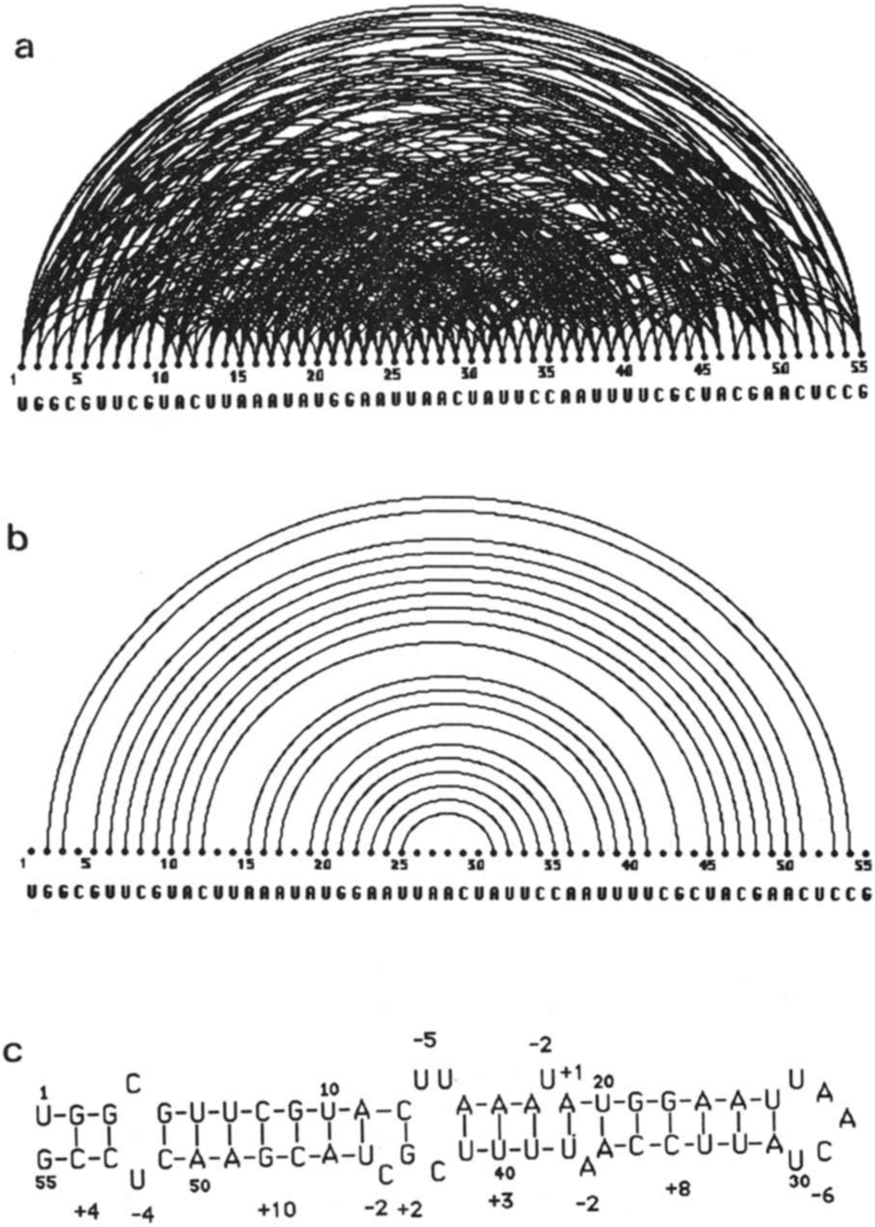

Fig. 4. (a) State of the system after the first iteration step.
(b) State of the system after the 61st iteration step.
(c) Secondary structure predicted by our algorithm.

has recently been developed (Zaritsky and Forester, 1991), to aid in the analysis of the partial RNase digestion data by displaying ideal autoradiograms. The 5′-terminal 75 bases of murine J-chain mRNA with a high potential for forming secondary structures with presumed biological activities (Zaritsky et al., 1988; Ben-Alon, 1989), served as an example to demonstrate the applicability of GEL. The program simulates the pattern of migration on an electrophoretic gel of a set of denatured RNA fragments obtained by partial cleavage of the native structure by each of the selected enzymes. However, GEL can only predict experimental results if some structure is assumed, a structure that is obtained by modelling as discussed above (Section 1). If one of GELs predictions is discordant with the corresponding experimental results (e.g., Knapp, 1989), the tested structure should be abandoned and another one may be introduced and tested; the procedure can be repeated until a particular predicted structure is confirmed.

GEL has two major disadvantages: (a) it can only confirm the existence of such a structure, not predict one itself; (b) the procedure to confirm one particular folding may be exceedingly time-consuming, because it depends on much luck in the order of choosing among the many possible alternative structures, predicted by theoretical considerations.

4.3. The direct, 'forward' approach

Instead of using the computer as a simulator for the autoradiogram from a hypothetical structure (Section 5) obtained by other procedures (1) as does GEL (Zaritsky and Forester, 1991), one can better exploit computer's potential in a more sophisticated way by having it analyze the data obtained from scanning an experimental autoradiogram and deduce the extant structure, rather than predict the autoradiogram.

4.3.1. Scanning procedures

Processing of the autoradiogram is done in two phases. In the first, the autoradiogram is scanned and each lane is compared against the ladder lane. This results within a set of points, each with its single strand characteristics (and double strand, in native conditions only). This phase uses standard line-detection techniques to identify the RNA fragments and filter out noise lines (Rosenfeld and Kak, 1982). Similar techniques are used for DNA autoradiograms (Gray et al., 1984; Russell et al., 1984; Elder et al., 1986; West, 1988). The identification can be done by either processing each of the lanes separately or by analyzing them as one lane. The above process is repeated for several specific RNases so that as much information as possible is gathered. This results in a set of points along with the following attributes: (a) its distance from the beginning of the sequence (the base's number); (b) the enzyme which 'discovered' the point and its 'coupling' characteristics.

The final outcome of this stage is represented by a base sequence of the relevant RNA molecule (first attribute), with the paired bases (second attribute) marked specifically (see e.g. the *M* format of Fig. 2 in Zaritsky and Forester, 1991).

4.3.2. Deriving the structure

The aim of this stage is to transform the *M* format to a planar representation of the structure (e.g., the *H* (hairpin) format of Fig. 2 in Zaritsky and Forester, 1991; and see here Figs. 2a, d and 4c). This problem is analogous to the one solved for small RNAs by Takefuji et al. (1990b), and can therefore be incorporated within a neural network structure (Section 1.1) and handled by the MIS algorithm (Section 2) or the maximum planar subgraph algorithm (Section 3). The procedure of marking bases reduces the number of pairings and thus the number of bases involved in the structure, hence reducing the number of required processing elements in both algorithms. Consequently, it decreases the computation time. Information on the marked bases also provides an accurate prediction of the structure by reducing the number of valid alternatives: the chance to obtain the global optimum solution among many local minima is enhanced due to the smaller number of possible base pairings.

Ideally and as a first approximation, all the marked bases in the molecule under study are paired and the only paired bases are those marked; in other words, there is a unique correspondence between a mark and a pairing. (As a corollary, there should be an even number of marked bases in the resultant *M* format. Cases in which this condition is not fulfilled will be discussed briefly below in Section 5.)

5. Additional problems

One can expect several kinds of problems in this approach, because the experimental procedures are subject to various diversions from the ideal picture drawn (Section 1). The following are examples for which solutions can (a) or cannot (b) be easily incorporated: (a) secondary recognition sites for enzyme cleavages; (b) interactions of certain stretches of the studied RNA with any regulatory molecule or structure in the living cell (be it DNA, RNA, protein, low molecular weight molecules or a combination of these).

5.1. Secondary recognition sites

This artifact results in weak bands on the autoradiogram, which can be treated rigorously, as for the following example: RNase T1 usually cleaves downstream (at the 3′-end of) any unpaired G and does not cleave paired Gs. However, lower cleavage efficiencies are occasionally observed downstream when a paired G has a 3′ neighboring U which is not paired. The resultant weaker band can either be ignored by the image analysis procedure (Section 4.3.1), or it can be labeled as a second-type mark on the *M* format output (Zaritsky and Forester, 1991; Section 4.2) and recognized by a simple addition to the algorithm (Sections 2 and 3; unpublished).

5.2. Regulatory cellular molecules

Many in vivo regulatory signals involve interactions between macromolecules, low molecular weight molecules or structures. A pertinent example is the temporary binding of a ribosome to an mRNA, which is necessary for translation of the latter to a protein. The nature of this signal is not fully understood yet, and much effort is being expended to decipher this interaction, which is crucial for any living.cell (e.g., Kozak, 1986). Any stretch of RNA involved in such binding is protected from cleavage(s) by the specific RNase(s), thus eliminating the relevant band(s) from the autoradiograms, bands which usually appear in the in vitro reaction mixtures (not commonly containing ribosomes). Such an interaction will be overlooked because these bands will show up in the absence of ribosomes. The only way to avoid this difficulty is by performing simulated 'real life' experiments (e.g., adding ribosomes to the mixture).

Acknowledgments

Dr. Ehud Gudes, Eng. Edan Forester and Mrs. Monica Einav were helpful in various stages of this work.

References

Ben-Alon, D., 1989, Secondary structures in mRNA with biological activities. M.Sc. Thesis, Ben-Gurion University of the Negev.

Cimino, G.D., Gamper, H.B., Issacs, S.T. and Hearst, J.E., 1985, Psoralens as photoactive probes of nucleic acid structure and function: organic chemistry, photochemistry and biochemistry. Annu. Rev. Biochem. 54, 1151-1193.

Dahlberg, J.E. and Abelson, J.N., 1989, RNA processing (Part A: General Methods), Methods Enzymol. 180.

Dahlberg, J.E. and Abelson, J.N., 1990, RNA Processing (Part B: Specific Methods), Methods Enzymol. 181.

Diener, T.O., 1987, The Viroid (Plenum Press, New York).

Ehresmann, C., Bandin, F., Mongel, M., Roinby, P., Ebel, J.P. and Ehresmann, B., 1987, Probing the structure of RNAs in solution. Nucl. Acids Res. 15, 9109-9128.

Elder, J.K., Green, D.K. and Southern, E.M., 1986, Automatic reading of DNA sequencing gel autoradiographs using a large format digital scanner. Nucl. Acids Res. 14, 417-424.

Fresco, J.R., Alberts, B.M. and Doty, P., 1960, Some molecular details of the secondary structure of ribonucleic acid. Nature 188, 98-101.

Garrett, R.A., Douthwaite, S. and Noller, H.F., 1981, Structure and role of 5S RNA-protein complexes in protein synthesis. Trends Biochem. Sci. 6, 137-139.

Gerhart, E., Wagner, H. and Nordstrom, K., 1986, Structural analysis of an RNA molecule involved in replication control of plasmid R1. Nucl. Acids Res. 14, 2523-2538.

Goringer, H.U. and Wagner, R., 1988, 5S RNA structure and function. Methods Enzymol. 164, 721-747.

Gray, A.J., Beecher, D.E. and Olson, M.V., 1984, Computer based image analysis of one dimensional electrophoretic gels used for the separation of DNA restriction fragments. Nucl. Acids Res. 12, 473-491.

Gultyaev, A.P., 1991, The computer simulation of RNA folding involving pseudoknot formation. Nucl. Acids Res. 19, 2489-2494.

Holbrook, S.R., Sussman, J.L., Warrant, R.W. and Kim., S.-H., 1978, Crystal structures of yeast phenyl-alanine transfer RNA. J. Mol. Biol. 123, 631 - 660.

Holley, L.H. and Karplus, M., 1989, Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86, 152 - 156.

Jacobson, A.B., Kumar, H. and Zuker, M., 1985, Effect of spermidine on the conformation of bacteriophage MS2 RNA. J. Mol. Biol. 181, 517 - 531.

Jaeger, J.A., Turner, D.H. and Zuker, M., 1989, Improved predictions of secondary structures for RNA. Proc. Natl. Acad. Sci. U.S.A. 86, 7706 - 7710.

Karplus, M. and Petsko, G.A., 1990, Molecular dynamics simulations in biology. Science 247, 631 - 639.

Knapp, G., 1989, Enzymatic approaches to probing of RNA secondary and tertiary structure. Methods Enzymol. 180, 192 - 212.

Kozak, M., 1986, Influence of mRNA secondary structure on initiation by eucaryotic ribosomes. Proc. Natl. Acad. Sci. U.S.A. 83, 2850 - 2854.

Le S.-Y. and Zuker, M., 1991, Predicting common foldings of homologous RNAs. J. Biomol. Struct. Dynam. 8, 1027 - 1044.

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.

Mankin, A.S., Skripkin, E.A., Ckichkova, W.U., Kopylov, A.M. and Bogdanov, A.A., 1981, An enzymatic approach for localization of oligodeoxynucleotides binding sites on RNA. Fed. Eur. Biochem. Soc. Lett. 131, 253 - 256.

Martinez, H.M., 1988, An RNA secondary structure workbench. Nucl. Acids Res. 16, 1789 - 1798.

McCulloch, W.S. and Pitts, W.H., 1943, A logical calculus of ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115 - 133.

Mei, H.-Y., Kaaret, T.W. and Bruice, T.C., 1989, A computational approach to the mechanism of self-cleavage of hammerhead RNA. Proc. Natl. Acad. Sci. U.S.A. 86, 9727 - 9731.

Noller, H.F., 1984, Structure of ribosomal RNA. Annu. Rev. Biochem. 53, 119 - 162.

Nussinov, R., Pieczenik, G., Griggs, J.R. and Kleitman, D.J., 1978, Algorithm for loop matching. SIAM J. Appl. Math. 35, 68 - 82.

Pieler, T., Guddat, U., Oei, S.L. and Erdmann, V.A., 1986, Analysis of the RNA structural elements involved in the binding of the transcription factor IIIA from *Xenopus laevis*. Nucl. Acids Res. 14, 6313 - 6326.

Pipas, J.M. and McMahon, J.E., 1975, Method for predicting RNA secondary structure. Proc. Natl. Acad. Sci. U.S.A. 72, 2017 - 2021.

Puglisi, J.D., Wyatt, J.R. and Tinoco, I. Jr., 1991, RNA pseudoknots. Acc. Chem. Res. 24, 152 - 158.

Qian, N. and Sejnowski, T., 1988, Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865 - 884.

Reid, B.R., 1981, NMR studies on RNA structures and dynamics. Annu. Rev. Biochem. 50, 969 - 996.

Rosenfeld, A. and Kak, A.C., 1982, Digital Picture Processing (Academic Press, New York, N.Y.).

Russell, P.J., Crandall, R.E. and Feinbaum, R., 1984, Nucl. Acids Res. 12, 493 - 498.

Steeg, E.W., 1989, Neural network algorithms for RNA secondary structure prediction. Master's thesis, University of Toronto Computer Science Dept.

Stern, S., Weiser, B. and Moller, H.F., 1988, Model for the three-dimensional folding of 16S ribosomal RNA. J. Mol. Biol. 204, 447 - 481.

Takefuji, Y. and Lee, K.C., 1989, A near-optimum parallel planarization algorithm. Science 245, 1221 - 1223.

Takefuji, Y., Chen, L.L., Lee, K.C. and Huffman, J., 1990a, Parallel algorithms for finding a near-maximum independent set of a circle graph. IEEE Trans. Neural Networks 1, 263 - 267.

Takefuji, Y., Lin, C.W. and Lee, K.C., 1990b, A parallel algorithm for estimating the secondary structure in ribonucleic acids. Biol. Cybern. 63, 337 - 340.

Tinoco, I., Uhlenbeck, O.C. and Levine, M.D., 1971, Estimation of secondary structure in ribonucleic acids. Nature 230, 362 - 367.

Turner, D.H., Sugimoto, N., Jaeger, J.A., Longfellow, C.E., Freier, S.M. and Kierzek, R., 1987, Improved parameters for prediction of RNA structure. Cold Spring Harbor Symp. Quant. Biol. 52, 123 - 133.

West, J., 1988, Automated sequence reading and analysis. Nucl. Acids Res. 16, 1847 - 1856.

Williams, A.L., Jr. and Tinoco, I., Jr., 1986, A dynamic programming algorithm for finding alternative RNA secondary structures. Nucl. Acids Res. 14, 299 - 315.

Zaritsky, A., Gollop, R. and Cann, G.M., 1988, Tertiary structure of the mRNA coding for J-chain might be involved in differentiation of mature B-lymphocytes to $(IgM)_5$-secretory cells. Speculat. Sci. Technol. 11, 205 - 213.

Zaritsky, A. and Forester, E., 1991, A simulation program to display specific digestion products of predicted RNA foldings. Comput. Appl. BIOSci. 7, 57 - 60.

Zuker, M., 1989a, Computer prediction of RNA structure. Methods Enzymol. 180, 262 - 288.

Zuker, M., 1989b, On finding all suboptimal foldings of an RNA molecule. Science 244, 48 - 52.

Zuker, M. and Stiegler, P., 1981, Optimal folding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. 9, 133 - 148.