RNA Secondary Structure Prediction Using Soft Computing

Shubhra Sankar Ray and Sankar K. Pal

Abstract—Prediction of RNA structure is invaluable in creating new drugs and understanding genetic diseases. Several deterministic algorithms and soft computing-based techniques have been developed for more than a decade to determine the structure from a known RNA sequence. Soft computing gained importance with the need to get approximate solutions for RNA sequences by considering the issues related with kinetic effects, cotranscriptional folding, and estimation of certain energy parameters. A brief description of some of the soft computing-based techniques, developed for RNA secondary structure prediction, is presented along with their relevance. The basic concepts of RNA and its different structural elements like helix, bulge, hairpin loop, internal loop, and multiloop are described. These are followed by different methodologies, employing genetic algorithms, artificial neural networks, and fuzzy logic. The role of various metaheuristics, like simulated annealing, particle swarm optimization, ant colony optimization, and tabu search is also discussed. A relative comparison among different techniques, in predicting 12 known RNA secondary structures, is presented, as an example. Future challenging issues are then mentioned.

Index Terms—RNA, DNA, protein, combinatorial optimization, dynamic programming, soft computing, genetic algorithms, neural networks, fuzzy logic, metaheuristics, machine learning

1 Introduction

The availability of huge amount of biological data has opened a new direction in genomic analysis and structural prediction of DNA, RNA, and proteins in recent years. The challenge is to find an efficient way to use these rich trove of evidence to infer functional, biological, and structural properties. Many of these data, such as completely sequenced genomes, ribonucleic acids (RNAs), and proteins, have in turn led to an absolute demand for specialized tools to view, analyze, and predict the biological significance of the data. Throughout the last few decades, determining the RNA structure has gained significant attention of the researchers, as it is one of the key issues in understanding the genetic diseases and creating new drugs. It also helps the biologists to understand the role of the molecule in the cell [1], [2], [3], [4].

There are different types of RNA(s), e.g., transfer RNA (tRNA), messenger RNA (mRNA), viral RNA, ribozomal RNA, signal recognition particle RNA (SRP RNA). In a cell, the role of different RNAs varies widely from each other. For example, RNA is used as a genetic material, instead of DNA, by some viruses, and mRNA is used by all organisms in proteins synthesis. Small nuclear RNAs (snRNA) are those eukaryotic RNAs, which are involved in processing hnRNAs. Some RNAs are responsible for controlling gene expression, or sensing and communicating responses to cellular signals. Thus, RNAs play an important role for regulating many biochemical and signalling activities in cells [4].

The RNA secondary structure prediction problem is a critical one in molecular biology. Secondary structure as well as tertiary structure can be determined by x-ray crystallography [5], [6] and nuclear magnetic resonance (NMR) spectroscopy [7]. Techniques involving small-angle x-ray solution scattering (SAXS)[8], hydroxyl radical probing [9], [10], in-line probing [11], and modification of bases by selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) [12], dimethyl sulfate (DMS) [13], l-cyclohexyl-3-[2-morpholinoethyl]carbodiimide metho-p-toluene sulfonate (CMCT) [14], 1,1-dihydroxy-3-ethoxy-2-butanone (known as kethoxal) [15], nucleases [16], [17], diethyl pyrocarbonate [18], and ethylnitrosourea [18] are also successfully used to determine the secondary structure. In general, these processes are difficult, slow, and expensive. Moreover, most RNAs are currently impossible to crystallize. That is why developing mathematical and computational methods to predict the secondary structure of RNA is very necessary [19], [20], [21], [22], [23]. Data analysis tools used for prediction of RNA structure are mainly based on dynamic programming (DP) [19], [20], [21], [22], [24], [25], [26]. The role of soft computing tools, a collection of flexible information processing techniques [27], gained significance with the need to generate approximate and good solutions by considering the kinetic effects in RNA folding [28], [29], [30], cotranscriptional folding [29], [31], [32], [33], pseudoknots [34], and deficiencies in energy parameters [35]. The term "soft computing" refers to computing techniques those strive for soft decisions based on given tolerance of imprecision and uncertainty for a particular problem, instead of exactness, as strived by hard computing techniques. In this paper, we survey the role of various

- *S.S. Ray is with the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, 203, B.T. Road, Kolkata 700108, India. E-mail: shubhra@isical.ac.in.*
- *S.K. Pal is with the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, 203, B.T. Road, Kolkata 700108, India. E-mail: sankar@isical.ac.in.*

Manuscript received 2 Nov. 2011; revised 29 Nov. 2012; accepted 2 Dec. 2012; published online 10 Dec. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-11-0283. Digital Object Identifier no. 10.1109/TCBB.2012.159.

soft computing tools, such as genetic algorithms (GAs), artificial neural networks (ANNs), and Fuzzy logic (FL), for formulating different methodologies. The role of different metaheuristics, that are closely related with the GA in the iterative search process, such as simulated annealing (SA), particle swarm optimization (PSO), ant colony optimization (ACO), and tabu search (TS), is also discussed. While GAs are adaptive and robust search processes producing near optimal solutions and have a large amount of implicit parallelism in creating a number of solutions, ANNs are machinery for adaptation and curve fitting and FL deals with the concept of partial truth, approximate reasoning, and partial set membership for uncertainty handling. GAs are adaptive in the search process as they identify the good and bad solutions in the search space according to the fitness of the chromosomes, select a set of good solutions probabilistically in every iteration, and recombine partial solutions (through crossover operation) from the best chromosomes to form chromosomes of potentially higher fitness. For a neural network, curve fitting refers to the modeling of an approximate input output relationship, through supervised learning. Once the relation has been modeled, to the necessary accuracy, it can be used for structure prediction, with an input-output characteristic approximately equal to the relationship existing in the training set. For example, the curve fitting can be accomplished by training the neural network with vertex identification results of a tree like RNA structure [36].

The aim of the paper is to bring together the scattered research works in predicting RNA secondary structure under the different components of soft computing, and enable the researchers in both, biology and machine learning, to understand the relevant problems and issues of each other for furtherance of bioinformatics research. First, we describe the basic concepts in RNA structure prediction in Section 2. This includes the descriptions of helix, bulge, hairpin loop, internal loop, and multiloop. In Section 3, we explain some DP methods, as they are first used for prediction of RNA secondary structures. In Section 4, the relevance of different soft computing technologies in the said prediction task is discussed along with their characteristic features. Various methodologies developed for these purposes are described in Sections 5, 6, and 7. A comparative study among different soft computing tools, in predicting 12 known RNA secondary structures, is provided in Section 8 as an example. Some challenging research issues for enhancing computational intelligence of the methodologies are discussed in Section 9. Finally, conclusions and some future research directions are provided in Section 10, followed by an extensive bibliography. A very preliminary version of this review, mentioning the descriptions of only three existing methodologies, was reported in [37].

2 Basic Concepts in RNA Secondary Structure Prediction

Here, we introduce the biological concepts required to understand the RNA structure, and then we describe the various secondary structural elements. The effect of ions, temperature, and proteins on RNA is also provided.

2.1 Biological Basics in RNA Secondary Structure Prediction

An RNA molecule represents a long chain of monomers called *nucleotides*, and each nucleotide consists of a base (any one of adenine (A), cytosine (C), guanine (G), and uracil (U)), a phosphate group and a sugar group [3]. The specific sequence of the bases along the chain is called primary structure. The structure is usually defined over the alphabets "A," "C," "G," and "U." Through the formation of hydrogen bonds the two groups of complementary bases, A-U and C-G, form stable base pairs, known as the Watson-Crick base pairs [3], [38], [39]. While the A-U pairs form two hydrogen bonds, the C-G pairs form three hydrogen bonds and tend to be more stable than the A-U pairs. Other bases also sometimes pair, especially the G-U pair. The G-U pairs are known as wobble base pairs and form two hydrogen bonds [2], [40]. Note that different types (depending on strand orientation) of U-U, U-C, G-A, A-A, and A-C base pairs [41] and base triples like GAC, GGC, and so on are also found in RNA structures [42]. A base pair provides a groove for insertion of an unpaired nucleoside to form a triple. The base-pair structure is referred to as the secondary structure of RNA [24], [43], [44], [45]. Generally, the secondary structure is determined in terms of substructures by observing each base is either paired or not and structure formation for RNA molecules is dependent on the Watson-Crick base pairs, wobble base pairs, and stacking of adjacent base pairs.

2.2 Secondary Structural Elements in RNA

Seven recognized secondary structural elements exist in RNA, and these are:

1. Hairpin loop,
2. bulge loop,
3. internal loop,
4. multibranched loop,
5. single-stranded regions,
6. helix, and
7. pseudoknots.

A common subsubstructure also exists in many parts of RNA, as well as in hairpin loop, helices, and pseudoknots, called *stem*. When a stem is a part of a helix, we can refer it as a *helical stem*. A schematic view of various structural elements, within dotted boxes, is shown in Fig. 1. Stems also help to identify the start and end of most substructures. A four-letter alphabet is used to represent an RNA sequence, which is the primary structure of RNA. Let $S = s_1, s_2, \ldots, s_n$ be an RNA sequence, where base $s_i \epsilon$\{A, U, C, G\} and $1 \leq \mathrm{i} \leq \mathrm{n}$. The subsequence $s_{(i,j)} = s_i, s_i + 1, \ldots, s_j$ is a segment of S, where $1 \leq \mathrm{i} \leq \mathrm{j} \leq \mathrm{n}$. If s_i and $s_j \epsilon$\{A-U, C-G, U-G\}, then s_i and s_j may constitute a base pair (i, j). Each base can at most take part in one base pair. Now, we describe, in brief, different substructures within RNA secondary structure:

- *Single-stranded region*. Let r be a sequence of bases in an RNA sequence. If all the bases in r are not paired with any other bases in the RNA structure, then we say r is a single-stranded region.

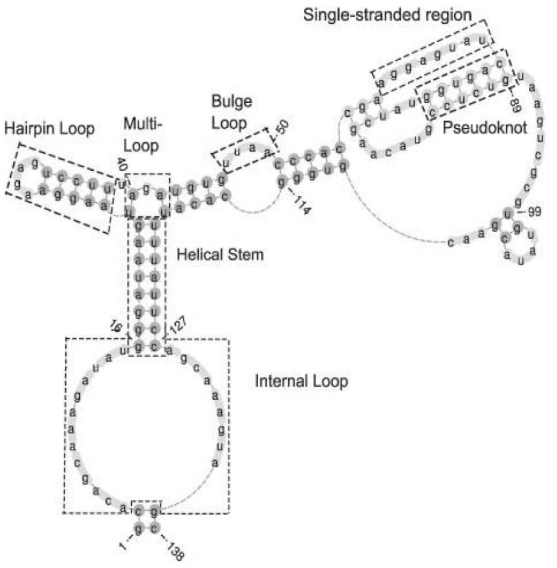

Fig. 1. Different types of secondary substructures in RNA.

- *Stem*. Contiguous stacked base pairs are called stems [46]. In this structure, base pairs are generally stacked onto other base pairs. If base pairs (i, j) and $(i+1, j-1)\epsilon S$ then, base pairs (i, j) and $(i+1, j-1)$ constitute the stack $(i, i+1 : j-1, j)$, and $m(\geq 2)$ consecutive stacks form the stem $(i, i+m : j-m, j)$ with the length of $m+1$.
- *Loop*. Single-stranded subsequence bounded by base pairs is called a loop [45].
- *Hairpin loop*. A loop at the end of a stem is called a hairpin loop. If $(i, j)\epsilon$ S but none of the elements, $i+1 \ldots j-1$, are paired with any other base, then the cycle is called a hairpin loop. Many molecular biologists use "hairpin" to refer to a stem with a loop of size 0 or 1 at the end, i.e., a stem with virtually no loop.
- *Bulge loop*. Single-stranded bases occurring within only one side of a stem are called a bulge loop. If substructure contains base pairs (i, j) and $(i+1, q)$, and there are some unpaired elements between q and j, then these unpaired elements form a bulge loop.
- *Internal loop*. In an internal loop, there are single-stranded bases interrupting both sides of a stem. If $i+1 < p < q < j-1$ and substructure contains base pairs (i, j) and (p, q), but the elements between i and p are unpaired and the elements between q and j are also unpaired, then the two unpaired regions constitute an internal loop.
- *Multibranched loop*. The loop from which three or more stems radiate is called a multibranched loop [47]. If substructure contains two or more base pairs like (i, j), (p, q), and (r, s), and the indices of none of the pairs lie within each other, then a multibranched loop is formed.
- *Helix*. In general, stems are considered to be helices [46], and they provide stability in the secondary structure. Single-stranded RNA folds back itself, forming helical areas interspersed with unpaired, single-stranded areas. Since the generation of a helix terminates at the first mismatched base pair, other secondary structures are implicitly defined in the various bulges and loops that remain outside of the stacked pairs. Thus, the determination of helices alone is considered sufficient, in some investigations [23], [48], [49], to account for all other secondary structure elements.
- *Pseudoknots*. Pseudoknots occur when unpaired bases of one substructure (e.g., the loop part in a hairpin loop) bind to unpaired bases of another substructure to form a stem [46]. If the resulting stem, formed from this type of bonding, stacks upon an existing stem, then the new formation is called coaxial stacking between two stems with a quasicontinuous helix. This structure also forms a pseudoknot and helps to stabilize the RNA structure. Even in pseudoknots, the major driving force of structure formation is Watson-Crick base pairs, Wobble base pairs, and stacking of adjacent base pairs.

In RNA, the primary sequence determines the secondary structure, which, in turn, determines its tertiary folding [1]. The secondary structural elements interact between themselves through formation of hydrogen bonds and Van der Waals interactions and fold in a 3D space to form the tertiary structure, a biologically active conformation. The folding process is facilitated by the presence and increasing concentration of cations (like magnesium ions), lowering temperature, and presence of proteins, called RNA chaperones [50], [51]. The interactions can take place between two helices, two unpaired regions, or one unpaired region and a double-stranded helix [50]. The effects of environment on RNA folding, involving ions, temperature, and proteins, are as follows:

- *Ions*. There are multiple sequential stages in folding process of RNA and different types of cations help in stabilizing these steps. In the early stages, secondary structure is formed and stabilized through the neutralization of the polyanionic backbone by binding with monovalent and divalent cations. In the later stages, tertiary structure is stabilized mostly through the binding of divalent cations such as magnesium [1]. Some studies suggest that monovalent and divalent cations compete with each other to stabilize the RNA tertiary structure [52]. However, the stabilizing is mainly achieved by binding of cations with backbone twisting sites, having high negative charge due to a high concentration of phosphates.
- *Temperature*. The effect of temperature on RNA can be understood by considering an RNA sequence, which is just unfolded in a solution at a specific high temperature [1]. Now, if the temperature is decreased slowly, some parts of the sequence begin to form stem and loops and the other parts will remain unfolded or partly folded. Some stems will grow longer with decreasing temperature, and some of them will be interrupted by internal loops and bulges. Finally, these stems, loops, and bulges will take part in the tertiary structure.
- *Proteins*. RNA chaperones are the proteins that can facilitate the RNA folding process. They work in

two ways, specific and nonspecific interactions. In specific interaction, the RNA folds around a given RNA binding platform, provided by the protein cofactor [51], and can stabilize a specific and thermodynamically favorable tertiary structure. After the formation of the stable structure, the protein is removed from the conformational equilibrium. In nonspecific interaction, the protein-assisted RNA folding leads to a stable structure, where the intermediate misfolded RNAs are destabilized and further misfolding is prevented. The overall process increases the probability of an RNA molecule to achieve its native structure. Note that, sometimes RNAs fold independently of proteins and then they interact.

Although, the tertiary structure is the level of organization relevant for biological function of structured RNA molecules and sometimes secondary structures are influenced by tertiary structures [53], the interactions, responsible for RNA tertiary structure formation, are significantly weaker than those responsible for secondary structure formation [1]. Hence, in recent investigations, for computational simplicity, it is assumed that the influence of hydrogen bonds within tertiary structure, on hydrogen bonds within secondary structure, is negligible; consequentially, secondary structures can be predicted independently of tertiary structures [35], [54], [55]. Current research is going on SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) [12], a powerful biochemical method for RNA structure probing, that uses reagents to modify the backbone in structurally flexible regions of RNA, and sequencing-based characterization of RNA structure [56]. Both SHAPE chemistry and sequencing-based techniques are primarily experimental procedures for structure elucidation rather than for structure prediction. While, the energy function (efn) in a prediction algorithm can be modified by adding pseudofree energy change terms, derived from SHAPE reactivities [12], the sequencing-based characterization data can be coupled with convex optimization-based prediction methods, through numerical algorithms [56], to improve the quality of predictions.

2.3 Example

Gibbs free energy ($\triangle G$) is generally used for calculating the total energy of different RNA structures (obtained from the same sequence) and the structure with minimum energy is accepted. This energy is a thermodynamic potential that measures the capacity of a system to do nonmechanical work. It is also the chemical potential that is minimized when a system reaches equilibrium at constant pressure and temperature. Hence, it is widely used to calculate fitness of the RNA secondary structure, obtained using various prediction algorithms [57]. It is represented as a function of enthalpy ($\triangle H$), temperature T, and entropy ($\triangle S$):

$$\triangle G = \triangle H - T\triangle S. \tag{1}$$

Enthalpy is a measure of the total energy of a thermodynamic system. The total energy is the sum of the internal energies. It is also the energy required to create a system, and the amount of energy required by the system for displacing its environment and establishing its volume and pressure. Temperature is a measure of the average kinetic energy in a system and entropy is a measure of how much of the energy of a system is potentially available to do work and how much of it will potentially manifest as heat. So, change in entropy can be used as a quantitative measure of the relative disorder of a system. Differences in Gibbs free energy ($\triangle G$), in a reaction or a conformation change, provide information on the process spontaneity. A positive free energy difference indicates that the reaction is in favor of the reactants and the reaction will go on spontaneously. A negative free energy difference in a reaction favors the products. A zero value of the free energy indicates that neither the reactants nor the products are favored.

Now, we will show how free energy can be calculated for a helix formation in an RNA secondary structure, using free energy thermodynamic parameters [58]. Let

$$\begin{pmatrix} 5'GAUC3' \\ 3'CUAG5' \end{pmatrix}$$

represent a helix within an RNA structure. The free energy change for this helix formation can be computed as

$$\begin{aligned} \triangle G(pred) = 2\triangle G\begin{pmatrix} 5'GA3' \\ 3'CU5' \end{pmatrix} + \triangle G\begin{pmatrix} 5'AU3' \\ 3'UA5' \end{pmatrix} \\ + \triangle G_{init} + \triangle G_{AUendpenalty}(perAUend) + \triangle G_{sym}. \end{aligned} \tag{2}$$

Using the values, downloaded from TURNER LAB, as mentioned in [58] and [59], the free energy becomes

$$\begin{aligned} \triangle G(pred) &= 2(-2.4) + (-1.1) + (4.09) + 0 + 0.43 \\ &= -1.38 \text{ kcal/mol}. \end{aligned} \tag{3}$$

In this example, the nearest neighbor terms are generated by considering a sliding window with two adjacent base pairs at the duplex RNA structure, and these results in three terms

$$\begin{pmatrix} 5'GA3' \\ 3'CU5' \end{pmatrix}, \begin{pmatrix} 5'AU3' \\ 3'UA5' \end{pmatrix}, \text{ and } \begin{pmatrix} 5'UC3' \\ 3'AG5' \end{pmatrix}.$$

The last term

$$\begin{pmatrix} 5'UC3' \\ 3'AG5' \end{pmatrix}$$

is similar to

$$\begin{pmatrix} 5'GA3' \\ 3'CU5' \end{pmatrix},$$

as one can be obtained by rotating the another by 180 degrees. There is a loss of entropy during initial pairing between the first two bases. This is accounted by a constant initiation term $\triangle G_{init}$. The term $\triangle G_{AUendpenalty}(perAUend) = 0$ as there are no AU base pair at the end for this helix. Furthermore, if there exists a GG mismatch (start of a loop) after the last AU pair then a bonus term is initiated. The last term $\triangle G_{sym}$ corrects for twofold rotational symmetry, resulting from self-complementary strand. If the two strands in the helix are not complementary then the last term is not considered for energy calculation. Moreover, if there is any GU base pair at the end of a helix, then the

$\triangle G_{GUendpenalty}$ should be taken into account. The free energy change for a hairpin loop with four or more unpaired nucleotides can be computed as

$$\begin{aligned}\triangle G(hairpin) = {} & \triangle G_{init}(n) + \triangle G_{terminalmismatch} \\ & + \triangle G_{GAorUUfirstmismatch} + \triangle G_{GGfirstmismatch} \\ & + \triangle G_{specialGUclosure} + \triangle G_{penalty}(allCloops),\end{aligned} \tag{4}$$

where n is the number of nucleotides in the loop part only, "terminal mismatch" parameter is the first mismatch stacking on the terminal base pair of a helix and not initiated for hairpin loop with three unpaired nucleotides, "GA or UU first mismatch" and "GG first mismatch" parameters are bonus terms, "special GU closure" is applicable to hairpins, where a GU closing pair (not UG) is preceded by two Gs, and the last term assigns a penalty for a loop with all C nucleotides. A tutorial for free energy calculation in various types of base pairs, helix, terminal mismatches, loops, and coaxial stacking is available at http://rna.urmc.rochester.edu/NNDB/turner04/index.html.

3 DP for RNA Secondary Structure Prediction

As dynamic programming algorithm (DPA) is first used to predict RNA secondary structure and is also used to predict substructures in some soft computing-based methodologies, it will be prudent to have an idea of this algorithm and how the research started in predicting RNA secondary structures. DP [19], [21], [24], [25], [26], [60] is a computational way to solve optimization problems, which can be divided into subproblems. At first, the optimal solution for each independent subproblem is calculated and then the solutions for overlapping subproblems are calculated by a recursive algorithm, repeatedly [60]. The optimal solutions for the subproblems are then preserved. The most probable secondary structure is predicted by calculating the sum of free energies, available from each optimal substructure, for all possible combinations of substructures and the combination with minimum free energy (MFE) is accepted.

One of the first attempts to predict RNA secondary structure, using DPA and by maximizing the number of base pairs using a simple nearest neighbor energy model, is presented in [60]. Nussinov and Jacobson [24] extended the method further. Zuker and Stiegler [20], [25] developed a popular DP-based algorithm, mfold, for finding the minimized free energy (MFE) pseudoknot-free secondary structure. In [60], an iterative definition of all secondary structures is first formulated and then functioned by minimizing the "distance" between segments of an RNA sequence, where "distance" is measured in terms of free energy. It is also assumed that the formation of a given base pair is independent of all other base pairs. The initial steps are based on the research work of Needleman and Wunch [61]. Then, the base pairing matrix, $p = (p_{i,j})$, for a given RNA sequence $s = s_1 s_2 \dots s_n$ (and the reversed order sequence $s' = s_n s_{n-1} \dots s_1$), is defined by $p_{i,j} = 1$ if s_i and s_j can form a bond, and $p_{i,j} = 0$ otherwise. When s_i and s_j are bonded, any bonding of $s_k (i < k < j)$ must be with points between i and j. The total number of structures having $i+1$ bonded pairs for a sequence $n+1$ long is given by a recursion relation. Let $N_{l,n}^{i}$ be the number of secondary structures, containing exactly i bonded pairs, formed on the subsequence $s_i s_{i+1} \dots s_n$. Then,

$$N_{l,n+1}^{i+1} = N_{l,n}^{i+1} + \sum_{j=l}^{n-m} \sum_{k=0}^{i} N_{l,j-1}^{k} N_{j+1,n}^{i-k} P_{j,n+1}, \tag{5}$$

where all hairpin loops have at least m bases. The equation follows from the fact that s_{n+1} is either bonded or not bonded. If s_{n+1} is not bonded, then there are $N_{l,n}^{i+1}$structures of interest. Otherwise, $n+1$ is bonded to some j, $l \le j \le n-m$, and if k bonds are formed in $s_l \dots s_{j-1}$, then $i-k$ must be formed in $s_{j+1} \dots s_n$. The definition of secondary structure implies that any combination of a k bonded structure with an $i-k$ bonded structure gives a secondary structure.

The mfold (multiple fold) webserver [20] uses the primary RNA sequence as input and predicts pseudoknot-free secondary structure with the MFE and some suboptimal structures. There are also options for users to choose the window for suboptimal structures, calculated in terms of percentage of the free energy of the MFE structure, and to force some selected base pairs in the energy calculation process to consider some auxiliary information into account. The core algorithm uses the DP method and provides the energy dot plot matrices for the base pairs contained within the foldings. In general, the mfold server provides an interactive medium to the user to select an window for suboptimal structures, certain base pairs and number of solutions at the output.

DP is also applied for calculating the equilibrium partition function for secondary structure, where the partition function is defined in terms of free energy, number of substructures, and temperature [62]. The partition function is used to calculate the probabilities of various substructures in terms of base pairs. The effect of partition function with increase in temperature, in unfolding transition of RNA, is also studied. The method provides an ensemble of secondary structures.

For a given RNA sequence, the software package Sfold (statistical fold) [63] computes the partition function for the ensemble of all possible secondary structures and draws samples according to their Boltzmann equilibrium probability, to form the Boltzmann ensemble [64]. Different clusters are then produced from it, and the centroid of the best cluster is chosen to provide the possible secondary structure [65].

The program RNAfold (RNA fold) from Vienna RNA package [22], [66] predicts a reliable secondary structure by checking the similarity between two structures obtained using DP under MFE model and centroid of the best cluster in Boltzmann ensemble. The Vienna RNA package also provides tools for RNA comparison, structural alignments, prediction of RNA-RNA interactions, and getting a clear perception about folding kinetics.

The KineFold [33] webserver can predict RNA structures with pseudoknots by considering the cotranscriptional folding on time scale. The method is based on addition or

removal of single helices and individual base-pair stacking/unstacking processes, as they are faster than nucleic acid folding and unfolding.

Though DPA traditionally yields optimal or suboptimal structures with MFE, it shows limitations in considering the kinetic effects related to easily accessible states in RNA folding [28], [29], [30], states with high energy barrier [30], [67], and conditions concerning suitability of transition from one folding state to another. Cotranscriptional folding is another unaddressed issue in DPA where the partial RNA sequence starts folding before the entire sequence has been transcribed [29], [31], [32], [33], [67]. These have prompted researchers to use soft computing tools that can handle kinetic effects in RNA folding, consider cotranscriptional folding, predict certain pseudoknots [34], and overcome some deficiencies in energy parameters [35] in a better fashion.

4 Relevance of Soft Computing in RNA Secondary Structure Prediction

Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capabilities for handling real-life problems. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low solution cost, and close resemblance with human-like decision making. The constituents of soft computing are mainly FL, ANNs, evolutionary algorithms (EAs), and rough sets (RS) [68]. The present paper mainly concerns with the application of GAs, ANNs, FL, and some metahuristics in RNA secondary structure prediction problem. The role of support vector machines (SVMs), which recently has gained attention of the researchers in pattern recognition (PR) and ANN is also discussed. At present, methods involving RSs, recurrent neural networks, radial basis function neural network, genetic programming, evolutionary strategies, and evolutionary programming are not available in the literature, for RNA secondary structure prediction. Although some of these methods are utilized for clustering or comparing the similarity of RNA secondary structures generated from various prediction algorithms, they are not structure prediction algorithms. Hence, these methods are not considered here.

As mentioned in Section 3, DPA-based methods try to find low-energy stable structures by neglecting the issues related with kinetic energy barriers, cotranscriptional folding, and so on. Studying every possible structure for a sequence would solve the RNA folding problem, but it is not always feasible. However, one may handle this problem in a soft computing paradigm where GAs can be used to navigate in the landscapes of structures and can provide a set of possible solutions and ANNs can be trained with models of known RNA structures and can predict a structure if it resembles with a previously trained model. FL can be used in conjunction with DPAs, GAs or metahuristics to adjust various parameters in the prediction process. On the other hand, multiple crossover and mutation operations, for a predefined population size, can be run in parallel using GAs to approximately solve the problem. Note that SVMs, a machine learning method, are used with an alternative approach, where RNA secondary structures can be predicted by comparative sequence analysis using functionally related sequences. However, soft computing-based methods are not deterministic and provide a number of suboptimal solutions. Therefore, one has to choose a consensus structure from the given solutions. In this regard, the interactive web interface in mfold server [20], Stem Trace visualizer [69], and analyzing histogram peak energy values for iteration-based methods may provide a direction to determine the consensus results. Although, the mfold server is designed to handle the DP results, the interface also allows a user to select base pairs, which can be chosen from the results of other methods.

Now, we discuss the characteristics of three soft computing methods, GAs [70], [71], [72], ANN [73], [74], and FL [75], that have been used in predicting the RNA secondary structure.

4.1 Characteristics of Different Soft Computing Methods

4.1.1 Characteristics of GAs

GAs [70], [71], [76] are adaptive heuristic search algorithms and premised on the evolutionary ideas of natural selection and genetics. The basic concept of GAs is to simulate processes in natural evolution that follow the principles of survival of the fittest. For optimization problems, GAs can provide robust, near-optimal, and fast solutions. They also have a large amount of implicit parallelism and provide a user-defined number (population of chromosome) of alternative close approximate solutions. Therefore, for many of the real-world problems, that involves finding optimal parameters and might prove difficult for traditional methods, the application of GAs appear to be an automatic choice. Since GAs show outstanding performance in optimization, it can be used for classification and clustering problems. GAs are proved to be useful and efficient when:

1. the search space is highly complex and large to perform an exhaustive search, and
2. conventional search methods cannot provide good solutions in a reasonable time.

GAs are executed with three basic operators: selection, crossover, and mutation. First, a binary or real-valued representation of possible solutions is coded such that it can be easily translated into a chromosome. Then, an appropriate objective function and associated fitness evaluation techniques are designed, taking the constraints into account. Finally, selection, crossover operation, and mutation operation are performed probabilistically and iteratively on a set of chromosomes, called *population*, to achieve near optimal solutions.

4.1.2 Characteristics of ANNs

ANN [73], [77] is a system composed of many simple processing elements (nodes) operating in parallel and is designed to emulate the biological neural network. ANNs are mainly used for function approximation, classification, prediction, feature extraction, and clustering. Depending on the task, the network can be supervised (SNN) or unsupervised (UNN). In SNN, the useful features within the data set are incorporated in the structure-activity

relationship model of the network by training it with known useful features. This, in turn, enables the network to detect correlations between the second- and higher-order patterns within data and finds the application of SNN in biological systems, showing nonlinear behavior. Using unsupervised neural networks (e.g., Kohonen self-organizing maps), tasks like feature extraction and data clustering, can be performed without knowing the class information of the data points. The main advantage of UNN is embedded in its unsupervised learning, where no previous knowledge about the data is required. The main characteristics of ANNs are:

1. it can easily adapt itself with the new patterns within data,
2. it encodes relation between the input and the output, however complicated, into network weights,
3. tolerance to distorted patterns (ability to generalize),
4. failure of components/nodes do not affect the performance of the system,
5. it can achieve high computational speed by using its components in parallel way, and
6. errors can be minimized through learning from examples (if input is A then output is B).

4.1.3 Characteristics of FL

FL deals with many-valued logic, where the reasoning is approximate rather than fixed and exact [75]. While in traditional logic theory, binary sets and crisp sets have two-valued logic, true or false, in FL the concept of partial truth is incorporated, which allows partial set membership, rather than crisp set membership. The variables or membership values in FL may have any value that ranges in between 0 and 1. It can also be implemented using nonnumeric linguistic variables such as low, medium, and high, where the variables may be managed by specific functions. The key idea is to relate the output with the inputs using if-statements. For example, if two bases are A and U, then the membership value for pairing is high (or 1.0), and if they are A and A then, the same may be low (or 0.3), depending on the application domain. Some important characteristics of FL are:

1. definite conclusions can be drawn from complex systems that generate vague, ambiguous, or imprecise information,
2. exact reasoning is viewed as a limiting case of approximate reasoning, and
3. any logical system can be fuzzified.

5 Soft Computing in RNA Secondary Structure Prediction

We now describe the existing techniques and methodologies developed for predicting RNA secondary structure, using different soft computing tools.

5.1 GAs for RNA Secondary Structure Prediction

Prediction of RNA secondary structure, using GAs, is investigated in [23], [28], [29], [48], [67], [78], [79], [80], [81]. While, some of the implementations use a binary representation for encoding the possible solutions (chromosomes in GAs) [28], [29], real coded representations, considering each substructure as separate integer, are also used as solutions in some investigations [23], [48]. In general, GAs generate conflicting stems (e.g., different stems sharing the same base pairs) that require removal of one of the conflicting stems by a repairing mechanism at later stage.

Van Batenburg et al. [28] developed a GA to predict RNA secondary structure, based on RNA folding pathways and free energy minimization. The method first creates a list of stems and a population of several possible solutions, each represented by a stem array. If a stem is present in a solution, then the corresponding position in the stem array is marked with 1, and otherwise 0. The method proceeds with standard selection, mutation, and crossover operations in binary GAs. The kinetic effects of RNA folding is incorporated by initially restricting the GA to a small part of the RNA sequence and then gradually increasing the sequence length by 10 percent of the initial string with each iteration of the GA. This resulted in inclusion of the whole RNA sequence in 10 iterations and temporary stems, which could be partially disrupted by another stem depending on the free energy values of the competing stems as well as on the loop that is formed when the stem is added. Finally, the algorithm is further improved by assigning different values for a "mutation that deletes a stem" and a "mutation that adds a stem" and squaring the fitness values of all stems to favor bigger stems. An improvement of this method is suggested in [29], where the kinetic character of stem formation and disruption is formulated by using probabilities, depending on the energy contribution of the stem. The concept of cotranscriptional folding is also incorporated by controlling the sequence length increment with the rate of energy improvement, obtained from the current GA iteration as compared to the previous one.

A massively parallel GA for RNA folding, based on free energy minimization and implemented in a computer with 16,384 processors and single instruction/multiple data (SIMD) architecture, is described in [78]. These processors are arranged in a 2D mesh with toroidal wraparound. At first, a stem pool is created using GA by considering each stem as a 4-tuple $(start, stop, size, energy)$, where b_i is the ith base in the RNA sequence, base $b_{start+i}$ pairs with base b_{stop-i} for $0 \leq i \leq size$, and "energy" represents the stacking energy of the stem. Stems are only generated for a user-given size or larger. Each processor then randomly selects a stem from the stem pool and goes on adding stems to complete a chromosome, which enables the GA to navigate in the landscape of structures. Finally, the secondary structure for each chromosome is represented by sorting these stems (with 4-tuples) w.r.t. the $start$ position and preserving it in a region table. For any two conflicting stems, the second stem is removed from the chromosome. Mutations are performed before the crossover operation by selecting a stem randomly from the stem pool and inserting it in a chromosome, and possible conflicts are handled in the crossover process. Uniform crossover operations are then performed by selecting parents, within two to eight neighboring processors and itself, using a ranked selection criterion. A stem is transferred from a parent to a child if there is no conflict

with already existing stems in the child. It is also pointed out that, in case of DPA, the selection of optimal substructures, emanating from a multibranch loop, may contribute more unpaired bases than global suboptimal solutions provided by GAs. An improved annealing-based mutation operator for this method is described in [79], where the total number of mutations drops linearly with generations of GA as the mutation probability descends hyperbolically with the size of the secondary structures.

The aforementioned parallel GA architecture for 16,384 processors is also used for predicting H-type pseudoknots [34], formed by interaction between a hairpin loop and a single-stranded region. Two stems are involved in this process by coaxial stacking and have no free nucleotides between them. On each processor, one list of stems and another list of H-type pseudoknots are maintained and initial secondary structures are formed from the list of stems. The H-type pseudoknots are then added to the initial structures at each generation of GA such that some of the new structures survive in the selection process of GA.

The concept of incorporating pseudostems in the stem pool of GA is introduced in [67] to accommodate multiple folding events and collision of stems in an RNA sequence. A pseudostem is a pair of two stems separated by an internal loop, across which coaxial stacking of base pairs occurs. Pseudostems ensure that some of the new structures will be fit enough to survive in the selection process of GA and help it to explore through the structures, which lies within the energy barrier for a traditional GAs and DPAs. It is also shown that while GA with high population size (e.g., 128k) predicts the rod-like linear structure, a lower population size (e.g., 4k) mostly predicts the metastable structures, and for intermediate population sizes the ratio of rod-like structures to metastable structures increases with population size. In a related investigation in [80], the accuracy of GA, fitness of chromosomes, and convergence time of GA, for various population sizes are studied for different data sets.

Wiese et al. [23] formulated the structure prediction task as a permutation of possible helices, using real coded GAs. At first, all potential helices are generated from a given RNA sequence by a helix generation algorithm [82], using a thermodynamic model. Each helix is then indexed with an integer ranging from 0 to $n-1$, where n is the total number of generated helices. Each chromosome of GA is encoded by a permutation of these integers and provides a solution for RNA secondary structure. In any chromosome, if two or more helices share some common base pairs, then for any two conflicting helices, the second helix is removed by checking all helices in a chromosome from left to right. Selection, crossover, and mutation operations are then applied to the chromosomes in a elitist model framework. At the end of the process, a set (population) of chromosomes with high scores, i.e., the chromosomes with minimized free energies, are accepted as possible solutions. Precautions are taken such that GAs not only optimize RNA structure in terms of MFE but also ensure that predicted structures are chemically feasible ones, i.e., any structure should not contain helices that share common bases.

The investigation in [23], is similar to that in [48], where a population of chromosomes evolves by selection, crossover, and mutation. The main difference between [48] and the recent method [23] has been the use of better crossover and mutation operators and incorporating state-of-the-art thermodynamic models to calculate the free energies. In [23], experimental results are provided by comparing the predicted structures with 19 known structures from four RNA classes.

GA is also utilized for estimating six free energy initiation parameters by maximizing the accuracy in predicting known RNA secondary structures [35]. While the first three initiation parameters are used in efn for standard DPA, the last three parameters are used in modified energy function (efn2) for a modified DPA to calculate the free energy in multibranch loops by considering coaxial stacking and number of unpaired nucleotides. The initial solutions of GA are generated from results obtained by optical melting of an RNA multibranch loop with three branching helices. In each step of GA, the six initiation parameters are randomly mutated and used to predict the RNA structure using DPA. The free energies of the predicted structures are scored and five sets of parameters for the top five structures are selected for next generation. The crossover operation is performed only in every fifth step of the GA by randomly selecting the first three parameters from one set and the last three parameters from another set to make a new set. Finally, the GA provides a set of initiation parameters, which enables the DPA to predict an RNA structure with highest accuracy.

In general, GAs provide a population of solutions as suboptimal structures and also make it possible to investigate not only the MFE structure but also other structures that may be closer to the natural fold. GAs seem suited to implementation for RNA structure prediction and also for estimating certain energy parameters.

5.2 ANN for RNA Secondary Structure Prediction

An RNA structure prediction method, involving graph-theoretic tree representation of RNA and training a three-layered back propagation ANN with vertex identification results of the tree, is presented in [36]. In the tree, stems are represented as edges, hairpins as vertex of degree one, internal loops and bulges as vertices of degree two, and junctions as vertices of degree three or more, using the technique of Le et al. [83]. The method is based on the hypothesis that tree representation of a secondary RNA structure can be achieved by merging the tree representation of smaller structures, out of many combinatorial possibilities. These include structures of known RNA, RNA-like candidates, and not RNA-like candidates. The vertex identification results for the $i\text{th}$ tree, resulting from the tree merge operation, are represented by a vector

$$p^i = \langle p_1^i, p_2^i, p_3^i, p_4^i \rangle, \tag{6}$$

where $i\text{th}$ tree is one of the training samples. Components of p^i correspond to 4 input nodes of a three-layered neural network with 24 nodes in the input layer and 2 nodes (say, y_1 and y_2) at the output layer. The neural network is trained to identify a tree as RNA or not. Here, $y_1 = 1$ and $y_2 = 0$

correspond to an RNA-like tree and $y_1 = 0$ and $y_2 = 1$ if the tree is not RNA like. The initial weights of hidden and output layer nodes are randomly assigned to values close to 0 and the error function is considered as

$$E = \frac{1}{2}\sum_{i=1}^{n} \|y(p^i) - q^i\|^2, \tag{7}$$

where $y(p^i) = \langle y_1(p^i), y_2(p^i)\rangle$ is the output corresponding to an input vector p^i, n is the total number of training samples (trees), $q^i = \langle 1, 0\rangle$ if the tree corresponds to known RNA or RNA-like structure, and $q^i = \langle 0, 1\rangle$ for not RNA-like structure. The weights are updated according to the standard procedure of back-propagation neural network and training continues until the error is close to 0. Finally, new tree structures are presented to the network for prediction task. The main advantage of this method lies in predicting new RNA structures, if they matches with known examples. A related investigation is available in [84].

In [85], the number of base pairs of RNA is maximized by using a Hopfield neural networks (HNNs) and circular graph representation. This representation was first introduced by Nussinov et al. [86], where the nucleotides are first aligned along the circumference of a circle graph and then the base pairs are represented by circular arcs that link paired bases. The number of neurons in HNN are considered to be the same as the number of base pairs, represented by the arcs of the circle. Each neuron is assigned the following binary function:

$$O_i = 1 \text{ if } I_i > 0, 0 \text{ otherwise},$$

where O_i and I_i are the output and input of the ith neuron, respectively. $O_i = 1$ indicates that the ith arc and the corresponding base pair are not included in the circle graph and vice versa. The neurodynamical model of ith McCullochPitts neuron is represented as:

$$\begin{aligned} dI_i/dt = {} & A(\Sigma_j^n d_{ij}(1 - O_j)(distance(i))^{-1})(1 - O_j)p(i)^{-1} \\ & - Bh(\Sigma_j^n d_{ij}(1 - O_j))O_i p(i), \end{aligned} \tag{8}$$

where $d_{xy} = 1$ if xth arc and the yth arc intersect each other in the circle graph, 0 otherwise, $h(x)$ is 1 if $x = 0$, 0 otherwise, A and B are the transfer functions and $p(i)$ is the absolute value of free energy of the ith base pair. A cost function, termed energy, is also introduced, where a neurons contribution to the energy is measured by the following equation:

$$\Delta E_k = E(a_k = 0) - E(a_k = 1) = (\Sigma_i a_i \omega_{ki}), \tag{9}$$

where a_k is the activation level of the ith neuron, and ω_{ki} is the connection weight between the ith and jth neuron. The state (on/off) of a neuron, in the network, is determined by the network itself and that state is selected, which lowers the networks energy.

In [87], class information of RNA in the initialization of Hopfield network is introduced. This resulted in improvement of experimental results with respect to the related investigation in [85].

5.3 FL for RNA Secondary Structure Prediction

A fuzzy DPA [88] is used to determine the RNA secondary structure in [89]. At first, multiple up-triangular matrices, each having combination of all possible base pairs, are constructed to store various substructures. Each element (i, j) in the up-triangular matrices is represented by a 4×4 matrix, mentioning the membership values for 16 possible base pair interactions (AA, AC, $\cdots$ UU). Fuzzy sets are separately used to partition the rows and the columns of the up-triangular matrices and known distributions of single bases from homologous RNA structures are incorporated as prior knowledge. For a particular position (i, j), the membership value is calculated by using the position specific membership information from all up-triangular matrices. The fuzzy DPA then iteratively updates the position information in all matrices and expands the base pair structure to predict the optimal structure such that, the product of the membership values, over all positions, are maximized.

6 Metaheuristics for RNA Secondary Structure Prediction

In this section, we discuss the role of different metaheuristics like SA, PSO, ACO, and TS, in RNA secondary structure prediction. Metaheuristics are closely related to GA, one of the components of soft computing, in the sense that they are computational method that optimizes a problem by iteratively trying to improve an initial solution with respect to a given measure of quality.

6.1 SA for RNA Secondary Structure Prediction

The use of SA [90] for predicting RNA secondary structures, using the free energy minimization approach, was first described by Schmitz and Steger [31]. The secondary structure is predicted by iterative formation and disruption of single base pairs through SA. Consequently, the energy changes are either changes in free energy or free activation energy. At the beginning, random formation and disruption of base pairs are allowed and the resulting unfavorable energy structures are subsequently suppressed by using a probabilistic selection process, based on Boltzmann factor. While the favorable structures are always accepted, the probability of accepting the new structure, with energy (E_{new}) greater than the old one (E_{old}), is computed by

$$Probability[Accept] < e^{-(E_{new} - E_{old})/R\theta}, \tag{10}$$

where temperature R is the Boltzmann's Constant and θ is the "distribution parameter," decreased gradually with each step of base pair selection method. The whole process is repeated for a predefined number of iterations or until θ achieves the desired value. The investigation also provides an idea of "sequential folding" during transcription by considering RNA polymerase chain elongation rates.

SARNA-Predict [49] employs a modified SA as its search engine combines a novel mutation operator and uses a free energy minimization-based approach. The method first creates initial solutions as a permutation of helices. New structures are then generated by mutating the existing ones and all new structures with reduced amount of free energy

are accepted. The mutation is accomplished by multiple swap operations between two randomly chosen helix positions. The process maximizes the chance of generating a new structure, which may not be achieved, using a single swap operation, due to the repairing process performed after mutation to obtain a valid RNA structure. The number of mutations is chosen as the product of the percentage of the total number of available helices and the current annealing temperature and, hence, varies with time. New structures with increased energy are also accepted with some probability, determined by the Boltzmann distribution, to avoid local minima in the search space. The probability of accepting a new structure with increased energy is computed by

$$Probability[Accept] = e^{-(E_{new}-E_{old})/T} = e^{-\Delta Cost/T}, \qquad (11)$$

where temperature T is the current temperature and E is the energy state. According to (11), the energy of a system at temperature T is probabilistically distributed among all different energy states in thermal equilibrium condition. In general, the process mostly accepts a downward step and sometimes accepts an upward step.

6.2 PSO for RNA Secondary Structure Prediction

A Set-based Particle Swarm Optimization algorithm (SetPSO), using mathematical sets to predict RNA structures with MFE, is described in [91]. For a given RNA sequence, all possible stems are first generated to form a universal set U. The secondary structure is then represented as a permutation of stems. Each stem is represented as a particle and the permutation of the particles is the vector representation of the PSO. In generic PSO, the position allocation of each particle (stem for RNA) and searching the best position can be performed by updating the position and velocity vectors, respectively, but in SetPSO, these are updated using two sets, instead of vectors. The first set is an open set, which contains elements that should be removed (subtracted) from the current position set. The second set is formed by adding a random subset of current position and a random subset of U. The traditional addition and subtraction operators are replaced, respectively, by the union and minus operations of the set. The personal best position of a particle is tracked by the particle itself during the update process and the final position set, a subset of U, provides a potential solution. Although it is indicated in [91] that DP-based mfold provides more accurate structures than SetPSO, according to citation information, it gained attention of the researchers in predicting RNA structures using PSO.

A generic PSO-based method, using fuzzy sets to adaptively adjust the weight, learning parameters, and particle number rate (PNR) in the velocity vector, is presented in [92]. The RNA structure is represented as a combination of stems and the free energy is minimized to predict the final structure. The globally best fitness (GOF) of the velocity vector and the number of generations for the best unchanged fitness (BUF) are considered as the inputs of the fuzzy system to adjust the aforementioned parameters, which in turn are used to update the velocity vector. While the ranges of the inputs, GOF and BUF, are mapped to [0, 1.0], in the output, those of weight, learning parameters and PNR, are set to [0.2, 1.2], [1.0, 2.0], and [0.5, 1.5], respectively.

6.3 ACO for RNA Secondary Structure Prediction

RNA secondary structure prediction problem, using ACO, is investigated in [93]. For a given RNA sequence, all possible stems are first identified using a brute-force algorithm and then new stems are probabilistically added by an ant to form a probable secondary structure, using ACO. The probability that an ant will select a stem is dependent on the previous stem and the pheromone trail between those two stems. The process is repeated for a number of ants to have multiple secondary structures, and the pheromone trails for all the structures are updated according to the best ant, in terms of minimized free energy, for a particular iteration. The pheromone trail acts as a memory for storing knowledge and provides the platform for learning in ACO, so that after multiple iterations, the algorithm provides some potential solutions for secondary structures. In [94], RNA folding pathways are simulated by calculating the energy of each stem and multibranch loop. Appropriate values for parameters are chosen from publicly available known secondary structures.

6.4 TS for RNA Secondary Structure Prediction

RNA secondary structure prediction based on TS, using the MFE model, is described in [95]. Stem, hairpin loop, internal loop, bulge loop, and multibranch loop are defined in terms of indices of the bases. For a particular iteration, a current solution is created with the longest stem, and other structural elements are then added to the current solution to generate neighboring solutions through intensification search, where a tabu list is used to avoid revisiting recently visited solutions. When all neighboring solutions, for an initial solution, are computed they are arranged in an ascending order of their free energy values. The initial solution is then updated with structural elements from the neighboring solutions. Finally, a diversification search is applied to explore the less frequently visited structural elements and to minimize the free energy of the current solution.

7 Other Methods

In this section, we discuss in brief about two machine learning techniques, k-nearest neighbor classifier and SVMs, those have gained attention of the researchers in RNA.

The classification process of a data point, using k-nearest neighbor algorithm (k-NN), is based on the majority voting among its k closest points in the feature space. Based on this classifier, the software package KnetFold (k-nearest neighbor classifiers-based Folding) predicts a secondary structure from alignment of multiple sequences. First, it predicts if any two columns of the alignment correspond to a base pair, using a voting scheme from the outputs of a hierarchical network of k-nearest neighbor classifiers, and then it generates a consensus probability matrix using RNAfold [22]. Finally, the last k-nearest neighbor classifier provides a consensus score by utilizing the consensus probability matrix value and the base pair probabilities from previous classifiers. The secondary structure corresponding to that score is then accepted as a solution.

TABLE 1
Description of RNA Sequences Taken from the Comparative RNA Website [102]

Organism	Accession Number	RNA class	Length	known bps
S.cerevisiae	X67579	5S rRNA	118	37
H. marismortui	AF034620	5S rRNA	122	38
M. anisopliae-2	AF197122	Group I intron, 23S rRNA	456	115
M. anisopliae-3	AF197120	Group I intron, 23S rRNA	394	120
A. lagunensis	U40258	Group I intron, 16S rRNA	468	113
H. rubra	L19345	Group I intron, 16S rRNA	543	138
A. griffini	U02540	Group I intron, 16S rRNA	556	131
C. elegans	X54252	16S rRNA	697	189
D. virilis	X05914	16S rRNA	784	233
X. laevis	M27605	16S rRNA	945	251
H. sapiens	J01415	16S rRNA	954	266
A. fulgens	Y08511	16S rRNA	964	265

SVMs, originally proposed by Vapnik [96], are a learning technique based on the statistical learning theory and has its genesis emerged from the principle of perceptron. SVMs, are used for mapping data points that cannot be separated by a linear hyperplane to a feature space so that the images of the data points in the feature space can be linearly separated. Considering a set of data points to be classified into two classes, a hyperplane is a generalization of the linear plane into a higher number of dimensions such that a good separation among data points is achieved by the hyperplane that has the largest distance to the training data points of any class. Most of the research works in predicting RNA secondary structure, using SVMs, are based on the alignment of RNA sequence with known sequence. In [97], RNA structure prediction problem is considered as a two-class classification problem and SVMs are used to predict whether two columns of sequence alignment form a base pair or not. The hyperplane in SVM is constructed by training it with positive and negative samples. The positive samples are those pair sites that form a base pair in the alignment of known sequences, and the negative samples are not. The feature vector for each pair site is composed of the covariation score, the base-pair fraction, and the base pair probability matrix. While the covariation score is a measure of complementary mutations, considering evolutionary information in the two columns of an alignment, the fraction of complementary nucleotides show the bias toward base pair for a pair of alignment columns. The base-pair probability matrix is a complementarity for detecting the conserved base pairs and the base-pair probability for every sequence in the alignment is computed with RNAfold [22]. These probability matrices are then aligned according to the sequence alignment and averaged. Considering the effect of sequence similarity upon covariation score, a similarity weight factor is also introduced, which adjusts the contribution of covariation and thermodynamic information toward prediction, based on sequence similarity. Finally, the common secondary structure is assembled by stem combing rules. The effectiveness and superiority of this method to related methods such as KnetFold [30], Pfold [98], and RNAalifold [99] are shown on 49 alignments. The method can also predict simple pseudoknots. Related investigations are available in [100], [101].

8 Comparison between Different Methods

Here, we compare the relative performance of some methods in predicting structures of 12 RNA sequences. These sequences are used in [49] and [23], and they represent 12 different organisms, different sequence lengths, and four different classes of RNA. The sequences are available in Comparative RNA website [102] and details are provided in Table 1. The quality of a predicted RNA structure, from a given sequence, can be judged either by the number of accurately predicted base pairs or by the level of the minimized free energy, and the number of true-positive (TP) base pairs and sensitivity can used as the RNA structure evaluation criterion and the performance evaluation criterion of the methods. While the value of TP base pairs for a given RNA sequence is the number of correctly identified base pairs among all predicted pairs, false positives are those non-base-pairs, which are incorrectly identified as base pairs. True negatives are non-base-pairs correctly identified as non-base-pairs and false negatives are those base pairs, which are incorrectly identified as non base pairs. The sensitivity is defined as

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}. \quad (12)$$

Considering the availability of results on the same RNA sequences and ease of implementation, a comparison among GA, SA, HNN, and mfold [20] in terms of predicted base pairs (predicted bps), TP base pairs, and sensitivity is provided in Table 2. The performance results for GA, SA, and mfold are taken from [23] and [49]. The HNN is implemented in a similar way as it is mentioned in [85] and the results are reported. While the GA and SA are two stochastic search methods and sometimes higher energy of an RNA structure is also accepted, the HNN and mfold are deterministic search methods and they move to a state to minimize the energy. The GA deals with a list of helix and a population of chromosomes, where each chromosome represents a possible solution and is represented by a helix array. The length of a chromosome is determined by the number of possible helices in the RNA sequence. The chromosomes then go through crossover and mutation operations, probabilistically, and the chromosomes are repaired to avoid helix repetition and conflict. The

TABLE 2
Comparison between Soft Computing Methods for Different RNA Secondary Structure Using Individual Nearest Neighbor with Hydrogen Bonds (INN-HB) Model

Sequence	Known bps	Predicted bps				TP				Sensitivity (%)			
		SA	GA	HNN	mfold	SA	GA	HNN	mfold	SA	GA	HNN	mfold
S.cerevisiae	37	39	39	38	41	33	33	19	33	89.2	89.2	50	89.2
H. marismortui	38	30	30	41	34	27	27	14	29	71.1	71.1	38	76.3
M. anisopliae-2	115	131	135	126	133	57	55	32	52	49.6	47.8	28	45.2
M. anisopliae-3	120	121	121	129	116	75	75	43	92	62.5	62.5	36	76.7
A. lagunensis	113	132	131	123	133	73	68	36	74	64.6	60.2	32	65.5
H. rubra	138	162	161	150	167	79	79	43	83	57.2	57.2	31	60.1
A. griffini	131	168	161	141	174	87	81	45	95	66.4	61.8	34	72.5
C. elegans	189	205	202	222	217	49	55	32	40	25.9	29.1	17	21.2
D. virilis	233	239	242	276	252	80	65	37	82	34.3	27.9	16	35.2
X. laevis	251	253	240	251	245	112	93	55	113	44.6	37.1	22	45.0
H. sapiens	266	244	250	266	258	116	89	53	95	43.6	33.5	20	35.7
A. fulgens	265	252	242	265	241	93	82	48	74	35.1	30.9	18	27.9
Averages	158.0	164.7	162.8	169.1	167.6	73.4	66.8	38.1	71.8	53.7	50.7	28.5	54.2

representation of an individual in SA may be similar to GA, but SA handles only one possible solution and, hence, mutation is the only possible operation. However, through the selection process, both the methods accept some solutions with lower fitness to avoid the problem of local minimum. On the other hand, the HNN deals with only one possible solution, and each neuron acts as a binary threshold unit and represents a base pair. The state (0 or 1) of a neuron indicates whether a base pair is present in a folding and the state is determined by a scalar value called "network energy." Finally, the HNN provides a possible solution through the minimization of network energy. The package mfold is based on DP, which deterministically and recursively searches for lower-energy structures, first in various parts of the sequence and then targets to integrate them to find the global minimum structure for the whole sequence. The method can provide optimal and suboptimal solutions with some restrictions to the size of loops. From the results of these methods in Table 2, we observe that the average results for TP base pairs and sensitivity are comparable for GA, SA, and mfold and their performances are superior to HNN. It is also found that SA performs a little better than GA and mfold when the number of known base pairs exceeds 180. mfold performs the best when the number of known base pairs is less than 150.

Figs. 2 and 3 compare the relative performances of GAs, SA, HNN, and mfold in terms of sensitivity versus the number of known base pairs and TP versus the number of known base pairs, respectively, for 12 different RNA sequences. From the figures, it is found that the curves for SA, GAs, and mfold are comparable and are at the top of the figures, whereas the curve for HNN lies below all other curves.

Now, we discuss in brief the time complexities of various algorithms that are used for comparison. In the GA, considering checking procedure for compatible helices, fitness calculation, one point crossover, and one point mutation, the time complexity is $O(g(xn + xb + xP_cn + xP_mn))$, where g is the number of generations, x is the initial size of population, b is the number of nucleotides, n is the number of possible helices, and P_c and P_n are crossover and mutation probabilities. Therefore, the asymptotic time complexity of the GA is $O(gxb)$, where $b > n$. The space complexity of GA is $O(xn)$ as it has to save the population. In the SA, considering multipoint swap mutation checking procedure for compatible helices and fitness calculation, the time complexity is $O(T(n + n + b))$, where T is the number of time steps, n is the number of possible helices, and b is the number of nucleotides. Hence, the asymptotic time complexity of the SA is $O(Tb)$, where $b > 2n$. The space complexity of SA is $O(b)$, as it has to preserve two structures in every step. Note that the initial structures of GA and SA are generated from the pool of helices. The time complexity of generating all possible helices is $O(n^2)$, and it is generally lower than the time complexities of GA and SA, mentioned before. The time complexity of HNN is $O(TN(N + N))$,

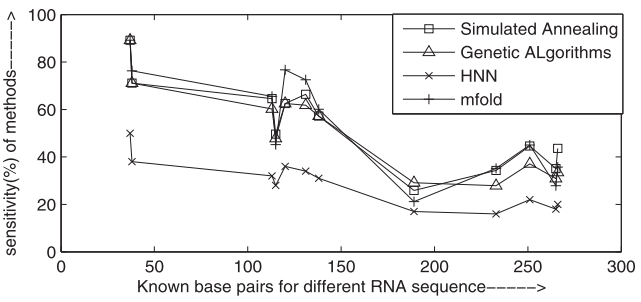

Fig. 2. Comparison among different methods in terms of sensitivity versus the number of known base pairs, for 12 different RNA sequences.

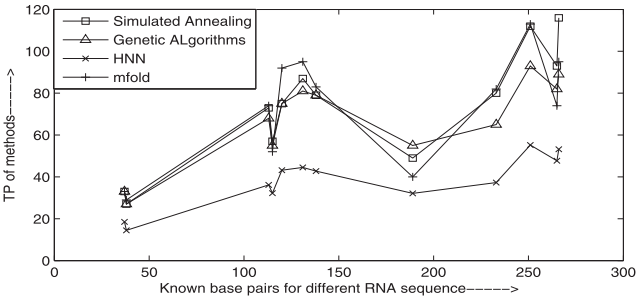

Fig. 3. Comparison among different methods in terms of TP base pairs versus the number of known base pairs, for 12 different RNA sequences.

where T is the number of iterations and N is the number of arcs in the circle graph. Therefore, the asymptotic time complexity is $O(TN^2)$. The space complexity of HNN is $O(n^2)$, where n is the length of the RNA sequence. The time complexity of the DPA used in mfold is $O(n^3)$ by fixing an upper limit to the size of an interior or bulge loop size to 30 nucleotides, where n is the number of nucleotides in the sequence. Without such a limit, the complexity is $O(n^n)$.

9 Challenging Issues

The 3D structure of RNA is a complex one and presents a new set of computational challenges to the bioinformatics community. The ability to relate the structural properties with the functional properties [103] may provide the key in solving these challenges. The different approaches for predicting RNA structure, discussed so far, involve only the efforts of some soft computing tools in their individual capacity. One of the major challenges in soft computing, namely, the symbiotic integration of its components, is still not yet addressed in the existing RNA literature. In this regard, integrated tools like neuro-fuzzy, rough evolutionary network, rough-fuzzy evolutionary network, and rough-fuzzy computing [104], may provide new directions in increasing the tractability in the application domain. It may also be mentioned here that fuzzy set theoretic models try to mimic human reasoning and can provide decisions having close resemblance with that of human.

Granular computing (GrC) has been proven to be a useful paradigm in mining data set, large in both size and dimension. When a problem involves incomplete, uncertain, and vague information, it may be difficult to differentiate distinct elements, and one may find it convenient to consider the data as granules, representing a group of data points that have very similar characteristics, and performing operations on them [105]. These characteristics can be obtained from similarity, equality, and closeness between the data points. For example, in neural networks, the self-organizing map (SOM) is a clustering technique that organizes the data in groups according to the underlying pattern. Each such group can be represented as an information granule. Incorporation of GrC in structure prediction task may also provide a conceptual framework for feature selection, classification, and clustering of the data. Here, fuzzy sets, RSs, and neural networks can be used in both, formulating granules and performing GrC.

Although the existing approaches for predicting RNA structure are useful, there is still some room for improving the output results. For example, in GAs, the basic crossover and mutation operators are common to all applications and can limit the effectiveness of GAs in structure prediction task; therefore, focused research to design more realistic and context sensitive operators is needed so that they can be coupled with the existing techniques. It should be mentioned here that GAs and metahuristics are more suitable than HNN for global optimization based tasks. So, investigations in maximizing the base pairs for RNA structure, using GAs or SAs rather than HNN, may provide encouraging results. The future research of RNA informatics will require integration of different soft computing tools in an efficient manner to enhance the computational intelligence in solving the related prediction problems, thereby signifying the collaboration between soft computing and RNA communities.

10 Conclusion

A review on some existing methodologies in soft computing framework for RNA secondary structure prediction problem is performed. In this regard, the basic concepts in RNA, different structural elements, and the effect of ions, proteins, and temperature on the RNA molecule are discussed. Brief descriptions of some DP-based software packages are also provided. The relevance of certain soft computing tools, especially GAs, are more explored. The comparisons among some existing methodologies, using 12 known RNA structures, revealed that average results are comparable for GA, mfold, and SA, and they are superior to HNN. Future challenging issues regarding the importance of relating the structural properties with the functional properties, integration of different soft computing methodologies, application of GAs, and different metahuristics in solving maximum independent set for prediction of RNA structure, and the need to design structure specific operators are addressed.

In some of the investigations, the hybridization of DP with FL, GAs, and metahuristics revealed a new research direction. First, substructures were predicted using DP and then the optimal or suboptimal structures were estimated using soft computing techniques. GAs appear to be a powerful soft computing tool to handle the task of structure prediction by not only considering the kinetic effects and cotranscriptional folding but also for estimation of certain free energy parameters. The existing metahuristics deal with permutation of substructures with MFE, but they have the potential to explore RNA folding pathways by itself creating the stems, pseudostems, and temporary stems, as performed by various GAs. They can also be utilized for predicting the pseudoknots and energy parameters. Although individual methods can compute the optimal or suboptimal structures within a given thermodynamic model, the natural fold of RNA is often in an energy state related to cotranscriptional folding or kinetic energy traps in folding landscape and requires soft computing-based methods to achieve those states.

Acknowledgments

The authors would like to thank anonymous reviewers for their suggestions in improving the quality of the review and Prof. Michael Zuker for providing the complexity of mfold algorithm. The work was done by Prof. S.K. Pal as a J.C. Bose Fellow of the Government of India.

References

[1] I. Tinoco Jr. and C. Bustamante, "How RNA Folds," *J. Molecular Biology*, vol. 293, no. 1, pp. 271-281, 1999.

[2] G. Varani and W.H. McClain, "The G x U Wobble Base Pair. A Fundamental Building Block of RNA Structure Crucial to RNA Function in Diverse Biological Systems," *EMBO Reports*, vol. 1, pp. 18-23, 2000.

[3] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, sixth ed. W.H. Freeman, 2007.

[4] E.A. Schultes and D.P. Bartel, "One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds," *Science,* vol. 289, pp. 448-452, 2000.

[5] S.H. Kim, G. Quigley, F.L. Suddath, and A. Rich, "High-Resolution X-Ray Diffraction Patterns of Crystalline Transfer RNA that Show Helical Regions," *Proc. Nat'l Academy of Sciences USA,* vol. 68, pp. 841-845, 1971.

[6] R.T. Batey, S.D. Gilbert, and R.K. Montange, "Structure of a Natural Guanine-Responsive Riboswitch Complexed with the Metabolite Hypoxanthine," *Nature,* vol. 432, pp. 411-415, 2004.

[7] A.E. Ferentz and G. Wagner, "NMR Spectroscopy: A Multifaceted Approach to Macromolecular Structure," *Quarterly Rev. of Biophysics,* vol. 33, pp. 29-65, 2000.

[8] R.P. Rambo and J.A. Tainer, "Improving Small-Angle X-Ray Scattering Data for Structural Analyses of the RNA World," *RNA,* vol. 16, pp. 638-46, 2010.

[9] R. Karaduman, P. Fabrizio, K. Hartmuth, H. Urlaub, and R. Luhrmann, "RNA Structure and RNA-Protein Interactions in Purified Yeast U6 snRNPs," *J. Molecular Biology,* vol. 356, pp. 1248-1262, 2006.

[10] T.D. Tullius and B.A. Dombroski, "Hydroxyl Radical 'Footprinting': High-Resolution Information about DNA-Protein Contacts and Application to Lambda Repressor and Cro Protein," *Proc. Nat'l Academy of Sciences USA,* vol. 83, pp. 5469-5473, 1986.

[11] E.E. Regulski and R.R. Breaker, "In-Line Probing Analysis of Riboswitches," *Methods Molecular Biology,* vol. 419, pp. 53-67, 2008.

[12] J.T. Low and K.M. Weeks, "Shape-Directed RNA Secondary Structure Prediction," *Methods,* vol. 52, no. 2, pp. 150-158, 2010.

[13] P. Tijerina, S. Mohr, and R. Russell, "DMS Footprinting of Structured RNAs and RNA-Protein Complexes," *Nature Protocols,* vol. 2, pp. 2608-2623, 2007.

[14] A. Bakin and J. Ofengand, "Four Newly Located Pseudouridylate Residues in Escherichia coli 23S Ribosomal RNA Are All at the Peptidyltransferase Center: Analysis by the Application of a New Sequencing Technique," *Biochemistry,* vol. 32, pp. 9754-9762, 1993.

[15] H.F. Noller and J.B. Chaires, "Functional Modification of 16S Ribosomal RNA by Kethoxal," *Proc. Nat'l Academy of Sciences USA,* vol. 69, pp. 3115-3118, 1972.

[16] R. Daou-Chabo and C. Condon, "RNase J1 Endonuclease Activity as a Probe of RNA Secondary Structure," *RNA,* vol. 15, pp. 1417-1425, 2009.

[17] P.W. Huber, "Chemical Nucleases: Their Use in Studying RNA Structure and RNA-Protein Interactions," *FASEB J.,* vol. 7, pp. 1367-1375, 1993.

[18] B. Singer, "All Oxygens in Nucleic Acids React with Carcinogenic Ethylating Agents," *Nature,* vol. 264, pp. 333-339, 1976.

[19] M. Zuker, "On Finding All Suboptimal Foldings of an RNA Molecule," *Science,* vol. 244, pp. 48-52, 1989.

[20] M. Zuker, "Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction," *Nucleic Acids Research,* vol. 31, pp. 3406-3415, 2003.

[21] M. Waterman and T. Smith, "Rapid Dynamic Programming Algorithms for RNA Secondary Structure," *Advances in Applied Math.,* vol. 7, no. 4, pp. 455-464, 1986.

[22] I.L. Hofacker, "Vienna RNA Secondary Structure Server," *Nucleic Acids Research,* vol. 31, pp. 3429-3431, 2003.

[23] K.C. Wiese, A.A. Deschnes, and A.G. Hendriks, "RnaPredict–An Evolutionary Algorithm for RNA Secondary Structure Prediction," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 5, no. 1, pp. 25-41, Jan.-Mar. 2008.

[24] R. Nussinov and A.B. Jacobson, "Fast Algorithm for Predicting the Secondary Structure of Single-Stranded RNA," *Proc. Nat'l Academy of Sciences USA,* vol. 77, no. 11, pp. 6309-6313, 1980.

[25] M. Zuker and P. Stiegler, "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information," *Nucleic Acids Research,* vol. 9, no. 1, pp. 133-148, 1981.

[26] T. Akutsu, "Dynamic Programming Algorithms for RNA Secondary Structure Prediction with Pseudoknots," *Discrete Applied Math.,* vol. 104, pp. 45-62, 2000.

[27] L.A. Zadeh, "Fuzzy Logic, Neural Networks, and Soft Computing," *Comm. ACM,* vol. 37, pp. 77-84, 1994.

[28] F.H. Van Batenburg, A.P. Gultyaev, and C.W. Pleij, "An APL-Programmed Genetic Algorithm for the Prediction of RNA Secondary Structure," *J. Theoretical Biology,* vol. 174, no. 3, pp. 269-280, 1995.

[29] A.P. Gultyaev, F.H. Van Batenburg, and C.W. Pleij, "The Computer Simulation of RNA Folding Pathways Using an Genetic Algorithm," *J. Molecular Biology,* vol. 250, pp. 37-51, 1995.

[30] E. Bindewald and B.A. Shapiro, "RNA Secondary Structure Prediction from Sequence Alignments Using a Network of K-Nearest Neighbor Classifiers," *RNA,* vol. 12, pp. 342-352, 2006.

[31] M. Schmitz and G. Steger, "Description of RNA Folding by Simulated Annealing," *J. Molecular Biology,* vol. 255, no. 1, pp. 254-66, 1996.

[32] I.M. Meyer and I. Miklós, "Co-Transcriptional Folding is Encoded within RNA Genes," *BMC Molecular Biology,* vol. 5, pp. 1-10, 2004, doi: 10.1186/1471-2199-5-10.

[33] A. Xayaphoummine, T. Bucher, and H. Isambert, "Kinefold Web Server for RNA/DNA Folding Path and Structure Prediction Including Pseudoknots and Knots," *Nucleic Acids Research,* vol. 33, pp. W605-W610, 2005.

[34] B.A. Shapiro and J.C. Wu, "Predicting H-Type Pseudoknots with the Massively Parallel Genetic Algorithm," *Computer Applications in the Biosciences,* vol. 13, pp. 459-471, 1997.

[35] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure," *J. Molecular Biology,* vol. 288, pp. 911-940, 1999.

[36] D.R. Koessler, D.J. Knisley, J. Knisley, and T. Haynes, "A Predictive Model for Secondary RNA Structure Using Graph Theory and a Neural Network," *BMC Bioinformatics,* vol. 11, pp. S6-S21, 2010.

[37] S.S. Ray, M. Bachhar, and S.K. Pal, "RNA Secondary Structure Prediction in Soft Computing Framework: A Review," *Proc. IEEE Third Int'l Conf. Computer Science and Information Technology,* vol. 5, pp. 430-435, 2010.

[38] J.D. Watson and F.H.C. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature,* vol. 171, pp. 737-738, 1953.

[39] A.L. Lehninger, D.L. Nelson, and M.M. Cox, *Principles of Biochemistry,* second ed. Worth, 1993.

[40] F.H. Crick, "Codon-Anticodon Pairing: The Wobble Hypothesis," *J. Molecular Biology,* vol. 19, pp. 548-555, 1966.

[41] N.B. Leontis and E. Westhof, "Geometric Nomenclature and Classification of RNA Base Pairs," *RNA,* vol. 7, pp. 499-512, 2001.

[42] E.A. Doherty, R.T. Batey, B. Masquida, and J.A. Doudna, "A Universal Mode of Helix Packing in RNA," *Nature Structural Biology,* vol. 8, pp. 339-343, 2001.

[43] D.H. Turner and N. Sugimoto, "RNA Structure Prediction," *Ann. Rev. of Biophysics and Biophysical Chemistry,* vol. 17, pp. 167-192, 1988.

[44] I. Tinoco, P.N. Borer, B. Dengler, M.D. Levin, O.C. Uhlenbeck, D.M. Crothers, and J. Bralla, "Improved Estimation of Secondary Structure in Ribonucleic Acids," *Nature New Biology,* vol. 246, pp. 40-41, 1973.

[45] R.W. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswick, and A. Zamir, "Structure of Ribonucleic Acid," *Science,* vol. 147, pp. 1462-1465, 1965.

[46] E. Westhof, B. Masquida, and F. Jossinet, "Predicting and Modeling RNA Architecture," *Cold Spring Harbor Perspectives in Biology,* vol. 3, pp. 1-12, 2011.

[47] N.B. Leontis and E. Westhof, "A Common Motif Organizes the Structure of Multi-Helix Loops in 16S and 23S Ribosomal RNAs," *J. Molecular Biology,* vol. 283, pp. 571-583, 1998.

[48] K.C. Wiese and E. Glen, "A Permutation-Based Genetic Algorithm for the RNA Folding Problem: A Critical Look at Selection Strategies, Crossover Operators, and Representation Issues," *Biosystems,* vol. 72, no. 1/2, pp. 29-41, 2003.

[49] H.H. Tsang and K.C. Wiese, "SARNA-Predict: Accuracy Improvement of RNA Secondary Structure Prediction Using Permutation Based Simulated Annealing," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 4, pp. 727-740, Oct.-Dec. 2010.

[50] E. Westhof and P. Auffinger, "RNA Tertiary Structure," *Encyclopedia of Analytical Chemistry,* R.A. Meyers ed., pp. 5222-32, John Wiley & Sons Ltd., 2000.

[51] M. Doetsch, R. Schroeder, and B. Furtig, "Transient RNA-Protein Interactions in RNA Folding," *The FEBS J.,* vol. 278, pp. 1634-1642, 2011.

[52] R. Shimana and D.E. Drapera, "Stabilization of RNA Tertiary Structure by Monovalent Cations," *J. Molecular Biology,* vol. 302, pp. 79-91, 2000.

[53] E. Koculi, S.S. Cho, R. Desai, D. Thirumalai, and S.A. Woodson, "Folding Path of P5abc RNA Involves Direct Coupling of Secondary and Tertiary Structures," *Nucleic Acids Research,* vol. 40, pp. 1-10, 2012.

[54] D.H. Mathews, "Predicting RNA Secondary Structure by Free Energy Minimization," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling,* vol. 116, pp. 160-168, 2006.

[55] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, and P. Schuster, "Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. II. Structures of Neutral Networks and Shape Space Covering," *Monatshefte für Chemie,* vol. 127, pp. 375-389, 1996.

[56] S. Aviran, C. Trapnell, J.B. Lucks, S.A. Mortimer, S. Luo, G.P. Schroth, J.A. Doudna, A.P. Arkin, and L. Pachter, "Modeling and Automation of Sequencing-Based Characterization of RNA Structure," *Proc. Nat'l Academy of Sciences USA,* vol. 108, no. 27, pp. 11069-11074, 2011.

[57] K.J. Doshi, J.J. Cannone, C.W. Cobaugh, and R.R. Gutell, "Evaluation of the Suitability of Free-Energy Minimization Using Nearest-Neighbor Energy Parameters for RNA Secondary Structure Prediction," *BMC Bioinformatics,* vol. 5, article 105, pp. 1-22, 2004.

[58] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner, "Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 7287-7292, 2004.

[59] D.H. Turner and D.H. Mathews, "NNDB: The Nearest Neighbor Parameter Database for Predicting Stability of Nucleic Acid Secondary Structure," *Nucleic Acids Research,* vol. 38, pp. D280-D282, 2009.

[60] M.S. Waterman and T.F. Smith, "RNA Secondary Structure: A Complete Mathematical Analysis," *Math. Biosciences,* vol. 42, pp. 257-266, 1978.

[61] S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Molecular Biology,* vol. 48, no. 3, pp. 443-453, 1970.

[62] J.S. McCaskill, "The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure," *Biopolymers,* vol. 29, pp. 1105-1119, 1990.

[63] Y. Ding, C.Y. Chan, and C.E. Lawrence, "Sfold Web Server for Statistical Folding and Rational Design of Nucleic Acids," *Nucleic Acids Research,* vol. 32, pp. W135-W141, 2004.

[64] Y. Ding and C.E. Lawrence, "A Statistical Sampling Algorithm for RNA Secondary Structure Prediction," *Nucleic Acids Research,* vol. 31, pp. 7280-7301, 2003.

[65] Y. Ding, C.Y. Chan, and C.E. Lawrence, "RNA Secondary Structure Prediction by Centroids in a Boltzmann Weighted Ensemble," *RNA,* vol. 11, no. 8, pp. 1157-1166, 2005.

[66] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast Folding and Comparison of RNA Secondary Structures," *Monatshefte für Chemie,* vol. 125, pp. 167-188, 1994.

[67] B.A. Shapiro, D. Bengali, W. Kasprzak, and J.C. Wu, "RNA Folding Pathway Functional Intermediates: Their Prediction and Analysis," *J. Molecular Biology,* vol. 312, pp. 27-44, 2001.

[68] S.K. Pal and A. Ghosh, "Soft Computing Data Mining," *Information Sciences,* vol. 163, pp. 1-3, 2004.

[69] W. Kasprzak and B.A. Shapiro, "Stem Trace: An Interactive Visual Tool for Comparative RNA Structure Analysis," *Bioinformatics,* vol. 15, pp. 16-31, 1999.

[70] D. Goldberg, *Genetic Algorithms in Optimization, Search, and Machine Learning.* Addison Wesley, 1989.

[71] L.B. Booker, D.E. Goldberg, and J.H. Holland, "Classifier Systems and Genetic Algorithms," *Artificial Intelligence,* vol. 40, nos. 1-3, pp. 235-282, 1989.

[72] S.K. Pal, S. Bandyopadhyay, and S.S. Ray, "Evolutionary Computation in Bioinformatics: A Review," *IEEE Trans. Systems, Man, and Cybernetics, Part-C,* vol. 36, no. 5, pp. 601-615, Sept. 2006.

[73] S. Haykin, *Neural Networks: A Comprehensive Foundation,* second ed. Prentice Hall, 1998.

[74] S.S. Ray, S. Bandyopadhyay, P. Mitra, and S.K. Pal, "Bioinformatics in Neurocomputing Framework," *IEE Proc. Circuits Devices and Systems,* vol. 152, no. 5. pp. 556-564, Oct. 2005.

[75] L.A. Zadeh, "Fuzzy Sets," *Information and Control,* vol. 8, pp. 338-353, 1965.

[76] M. Mitchell, S. Forrest, and J.H. Holland, "The Royal Road for Genetic Algorithms: Fitness Landscapes and GA Performance," *Proc. First European Conf. Artificial Life,* 1992.

[77] G.P. Zhang, "Neural Networks for Classification: A Survey," *IEEE Trans. Systems, Man and Cybernetics, Part C,* vol. 30, no. 4, pp. 451-462, Nov. 2000.

[78] B.A. Shapiro and J. Navetta, "A Massively Parallel Genetic Algorithm for RNA Secondary Structure Prediction," *J. Supercomputing,* vol. 8, pp. 195-207, 1994.

[79] B.A. Shapiro and J.C. Wu, "An Annealing Mutation Operator in the Genetic Algorithms for RNA Folding," *Computer Applications in the Biosciences,* vol. 12, pp. 171-180, 1996.

[80] B.A. Shapiro, J.C. Wu, D. Bengali, and M.J. Potts, "The Massively Parallel Genetic Algorithm for RNA Folding: MIMD Implementation and Population Variation," *Bioinformatics,* vol. 17, no. 2, pp. 137-148, 2001.

[81] K.C. Wiese, A.A. Deschnes, and E. Glen, "Permutation Based RNA Secondary Structure Prediction via a Genetic Algorithm," *Proc. Congress Evolutionary Computation,* pp. 335-342, 2003.

[82] K.C. Wiese and A. Hendriks, "Comparison of P-RnaPredict and Mfold-Algorithms for RNA Secondary Structure Prediction," *Bioinformatics,* vol. 22, no. 8, pp. 934-942, 2006.

[83] S. Le, R. Nussinov, and J. Maziel, "Tree Graphs of RNA Secondary Structures and Their Comparison," *Computers Biomedical Research,* vol. 22, pp. 461-473, 1989.

[84] T. Haynes, D. Knisley, and J. Knisley, "Using a Neural Network to Identify Secondary RNA Structures Quantified by Graphical Invariants," *Comm. Math. and Computer Chemistry/MATCH,* vol. 60, pp. 277-290, 2008.

[85] Q. Liu, X. Ye, and Y. Zhang, "A Hopfield Neural Network Based Algorithm for RNA Secondary Structure Prediction," *Proc. First Int'l Multi-Symp. Computer and Computational Sciences (IMSCCS '06),* pp. 1-7, 2006.

[86] R. Nussinov, G. Piecznik, J.R. Grigg, and D.J. Kleitman, "Algorithms for Loop Matching," *SIAM J. Applied Math.,* vol. 35, pp. 68-82, 1978.

[87] Q. Zou, T. Zhao, Y. Liu, and M. Guo, "Predicting RNA Secondary Structure Based on the Class Information and Hopfield Network," *Computers in Biology and Medicine,* vol. 39, no. 3, pp. 206-214, 2009.

[88] A.O. Esogbue and J. Kacprzyk, "Fuzzy Dynamic Programming: Main Developments and Applications," *Fuzzy Sets and Systems,* vol. 81, pp. 31-45, 1996.

[89] D. Song and Z. Deng, "A Fuzzy Dynamic Programming Approach to Predict RNA Secondary Structure," *Proc. Sixth Int'l Conf. Algorithms in Bioinformatics,* pp. 242-251, 2006.

[90] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, no. 4598, pp. 671-80, 1983.

[91] M. Neethling and A.P. Engelbrecht, "Determining RNA Secondary Structure Using Set-Based Particle Swarm Optimization," *Proc. IEEE Congress Evolutionary Computation,* pp. 6134-6141, 2006.

[92] C. Xing, G. Wang, Y. Wang, Y. Zhou, K. Wang, and L. Fan, "Psofold: A Metaheuristic for RNA Folding," *J. Computational Information Systems,* vol. 8, pp. 915-923, 2012.

[93] N. McMillan, "Rna Secondary Structure Prediction Using Ant Colony Optimisation," master's thesis, School of Informatics, Univ. of Edinburgh, pp. 1-63, 2006.

[94] J. Yu, C.H. Zhang, Y.N. Liu, and X. Li, "Simulating the Folding Pathway of RNA Secondary Structure Using the Modified Ant Colony Algorithm," *J. Bionic Eng.,* vol. 7, pp. 382-389, 2011.

[95] Y. Liu, J. Hao, and J. Peng, "Predicting RNA Secondary Structure with Tabu Search," *Proc. IEEE Int'l Conf. Cognitive Informatics,* pp. 409-414, 2010.

[96] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[97] Y. Zhao and Z. Wang, "Consensus RNA Secondary Structure Prediction Based on Support Vector Machine Classification," *Chinese J. Biotechnology,* vol. 24, no. 7, pp. 1140-1148, 2008.

[98] B. Knudsen and J. Hein, "Pfold: RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars," *Nucleic Acids Research,* vol. 31, pp. 3423-3428, 2003.

[99] S.H. Bernhart, I.L. Hofacker, S. Will, A.R. Gruber, and P.F. Stadler, "RNAalifold: Improved Consensus Structure Prediction for RNA Alignments," *BMC Bioinformatics,* vol. 9, article 474, pp. 1-13, 2008.

[100] Y. Sakakibara, K. Popendorf, N. Ogawa N, K. Asai, and K. Sato, "Stem Kernels for RNA Sequence Analyses," *J. Bioinformatics and Computational Biology,* vol. 5, pp. 1103-22, 2007.

[101] S. Qiu and T. Lane, "A Framework for Multiple Kernel Support Vector Regression and Its Applications to siRNA Efficacy Prediction," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 6, no. 2, pp. 190-199, Apr.-June 2009.

[102] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. DSouza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Muller, N. Pande, Z. Shang, N. Yu, and R. Gutell, "The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs," *BMC Bioinformatics,* vol. 3, no. 1, article 15, 2002.

[103] S.S. Ray, S. Halder, S. Kaypee, and D. Bhattacharyya, "HD-RNAS: An Automated Hierarchical Database of RNA Structures," *Frontiers in Genetics,* vol. 3, no. 59, pp. 1-10, 2012.

[104] *Rough-Fuzzy Hybridization: A New Trend in Decision Making,* S.K. Pal and A. Skowron, eds. Springer-Verlag, 1999.

[105] A. Ganivada, S.S. Ray, and S.K. Pal, "Fuzzy Rough Granular Self-Organizing Map and Fuzzy Rough Entropy," *Theoretical Computer Science,* vol. 466, pp. 37-63, 2012.

Shubhra Sankar Ray (www.isical.ac.in/~shubhra) received the BSc degree with physics honors, the MSc degree in electronic science, and the MTech degree in radio physics and electronics in 1998, 2000, and 2002, respectively, from the University of Calcutta, and the PhD degree in engineering in 2008 from Jadavpur University. He is an assistant professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He is also associated with the Center for Soft Computing Research, funded by the Department of Science and Technology, Government of India, at the same institute. He was a CSIR senior research fellow at Machine Intelligence Unit, Indian Statistical Institute in 2003-2006, visiting research fellow at Center for Soft Computing Research, Indian Statistical Institute in 2006-2008, and the postdoctoral fellow at the Biophysics Division, Saha Institute of Nuclear Physics, India, in 2008-2009. His current research activities are on soft computing, bioinformatics, neural networks, genetic algorithms, and data mining. Three of his journal publications are now curated in the *Saccharomyces* Genome Database. His name has been included in Marquis Who's Who in the World, 2007 and 2012 editions. He received the Microsoft Young Faculty Award 2010-2011 from the Microsoft Research Laboratory, India, and the Indian Statistical Institute.

Sankar K. Pal (www.isical.ac.in/~sankar) received the PhD degree in radio physics and electronics from the University of Calcutta, in 1979, and another PhD degree in electrical engineering along with DIC from Imperial College, University of London, in 1982. He is a distinguished scientist of the Indian Statistical Institute and its former director. He is also a J.C. Bose Fellow of the Government of India. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He joined his institute in 1975 as a CSIR senior research fellow where he became a full professor in 1987, distinguished scientist in 1998, and the director for the term 2005-2010. He was at the University of California, Berkeley, and the University of Maryland, College Park, in 1986-1987; the NASA Johnson Space Center, Houston, Texas, in 1990-1992 and 1994; and in US Naval Research Laboratory, Washington, DC, in 2004. Since 1997, he has been serving as a distinguished visitor of the IEEE Computer Society for the Asia-Pacific Region, and held several visiting positions in Hong Kong and Australian universities. He is a coauthor of 17 books and more than 300 research publications in the areas of pattern recognition and machine learning, image processing, data mining and web intelligence, soft computing, neural nets, genetic algorithms, fuzzy sets, RSs, and bioinformatics. He has received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India), and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, the 1998 Om Bhasin Award, the 1993 Jawaharlal Nehru Fellowship, the 2000 Khwarizmi International Award from the Islamic Republic of Iran, the 2000-2001 FICCI Award, the 1993 Vikram Sarabhai Research Award, the 1993 NASA Tech Brief Award, the 1994 *IEEE Transactions on Neural Networks'* Outstanding Paper Award, the 1995 NASA Patent Application Award, the 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, the 2005-2006 ISC-P.C. Mahalanobis Birth Centenary Award (Gold Medal) for Lifetime Achievement, the 2007 J.C. Bose Fellowship of the Government of India, and the 2008 Vigyan Ratna Award from Science and Culture Organization, West Bengal. He was also honored with Padma Shri in Science and Engineering 2013 (one of the highest civilian awards) by the president of India. He is/was an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2002-2006), *IEEE Transactions on Neural Networks* (1994-1998 and 2003-2006), *Neurocomputing* (1995-2005), *Pattern Recognition Letters*, *International Journal of Pattern Recognition and Artificial Intelligence*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, *Fundamenta Informaticae*, *LNCS Transactions on Rough Sets*, *International Journal of Computational Intelligence and Applications*, *IET Image Processing*, *Journal of Intelligent Information Systems*, and *Proceedings of the National Institute of Sciences of India;* editor-in-chief, *International Journal of Signal Processing*, *Image Processing and Pattern Recognition*; a Book Series Editor, *Frontiers in Artificial Intelligence and Applications*, IOS Press, and *Statistical Science and Interdisciplinary Research*, World Scientific; a member, Executive Advisory Editorial Board, *IEEE Transactions on Fuzzy Systems*, *International Journal on Image and Graphics*, and *International Journal of Approximate Reasoning*; and a guest editor of the *IEEE Computer*. He is a fellow of the IEEE, the Academy of Sciences for the Developing World (TWAS), Italy, International Association for Pattern Recognition, International Association of Fuzzy Systems, and all four National Academies for Science/Engineering in India.