Swarm and Evolutionary Computation 27 (2016) 68–79

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

journal homepage: www.elsevier.com/locate/swevo

Regular Paper

An efficient two-level swarm intelligence approach for RNA secondary structure prediction with bi-objective minimum free energy scores

Soniya Lalwani [a,c], Rajesh Kumar [b,*], Nilama Gupta [c]

[a] R&D, Advanced Bioinformatics Centre, Birla Institute of Scientific Research, Jaipur, India
[b] Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India
[c] Department of Mathematics, Malaviya National Institute of Technology, Jaipur, India

ARTICLE INFO

Article history:
Received 15 January 2015
Received in revised form
15 September 2015
Accepted 30 September 2015
Available online 19 October 2015

Keywords:
Particle swarm optimization
Minimum free energy
Hydrogen bond model
Nearest neighbor database
Combinatorial optimization
RNA secondary structure

ABSTRACT

This paper introduces a set-based two-level particle swarm optimization algorithm (TL-PSOfold) with multiple swarms for finding secondary structure of RNA with prediction accuracy. First objective is concerned with maximizing number of stacked loops at hydrogen bond, whereas, second objective deals with minimum free energy (MFE) at standard nearest neighbor database (NNDB). First level of the algorithm works on the entire search space for the best solution of each swarm, whereas, the second level works at the *gbest* solution of each swarm. The set based PSO approach has been applied at both levels to represent and update the set of ordered pairs of the folded RNA sequence. Improved weight parameter schemes with mutation operators are implemented for better convergence and to overcome the stagnation problem. Bi-objectives nature of TL-PSOfold enables the algorithm to achieve maximum matching pairs as well as optimum structure at respective levels. The performance of TL-PSOfold is compared with a family of PSO based algorithms i.e. HelixPSO v1, HelixPSO v2, PSOfold, SetPSO, IPSO, FPSO, popular secondary structure prediction software RNAfold, mfold and other metaheuristics RNA-Predict, SARNA-Predict at the criteria of sensitivity, specificity and *F*-measure. Simulation results for TL-PSOfold show that it yields higher prediction accuracy than all the compared approaches. The claim is supported by the non-parametric statistical significance testing using Kruskal–Wallis test followed by post-hoc analysis.

1. Introduction

Ribonucleic acid (RNA) is one of the three main biological macromolecules that are important for all known forms of life (other two being DNA and protein). It plays crucial role in protein synthesis, DNA replication, regulation and gene expression. Major classification of RNA for translation comprises four classes: messenger RNA (mRNA); transfer RNA (tRNA); ribosomal RNA (rRNA) and transfer-messenger RNA (tmRNA). The mRNA conveys information about a protein sequence to the ribosomes; tRNA transfers a specific amino acid at the ribosomal site of protein synthesis; ribosome of rRNA binds mRNA to execute protein synthesis; tmRNA tags particular kind of proteins. Besides these RNAs, several other RNAs too play active roles within cells [1].

Due to the key role of RNA in cell functioning, it becomes an indispensable task to find optimal structure of RNA for prediction of its function. Two types of approaches are available to determine structure of RNA: physical approaches and computational approaches. Physical approaches include X-ray crystallography [2] and NMR (nuclear magnetic resonance) spectroscopy [3]. The basic drawbacks of these approaches are their sensitiveness to experimental errors, dependence on environmental conditions, being very expensive and time consuming. Computational approaches are becoming very popular because secondary structure plays significant role in determining tertiary structure and molecule functioning. Most popular computational approaches are based on transformational grammar and thermodynamic parameters. One of the most popular example of transformational grammar based approach is Contrafold [4] based on conditional log-linear models, which generalize upon stochastic context-free grammars (SCFGs). The most popular thermodynamic parameters based approaches include RNAfold [5] and mfold [6], both based on dynamic programming. RNAfold from Vienna RNA package finds maximum similarity between two structures at minimum free energy (MFE) and the centroid of the best cluster in Boltzmann ensemble. The mfold also predicts sub-optimal structures. A comprehensive

[*] Corresponding author.
E-mail addresses: slalwani.math@gmail.com (S. Lalwani), rkumar.ee@gmail.com (R. Kumar), guptanilama@gmail.com (N. Gupta).

http://dx.doi.org/10.1016/j.swevo.2015.09.008
2210-6502/© 2015 Elsevier B.V. All rights reserved.

survey of all soft computing approaches developed so far for RNA secondary structure prediction has been performed by Ray and Pal (2013) [7].

There have been many recent reviews on particle swarm optimization (PSO) algorithm variants [8–11]. PSO variants for RNA secondary structure prediction include HelixPSO version 1 (HelixPSO v1) [12], SetPSO [13], HelixPSO version 2 (HelixPSO v2) [14], Improved PSO (IPSO) [15], Fuzzy adaptive PSO (FPSO) [16] and PSOfold [17].

HelixPSO v1 uses a multiple swarm approach, with the strategy for particles to have a target set of reference positions for defining the direction of movement of a particle. The starting positions of the particles are computed by genetic algorithm (GA). SetPSO adapts PSO to optimize the variable lengthen set-based problem of RNA helices. SetPSO successfully obtains optimal or near optimal solutions but not remarkably good solutions. HelixPSO v2 uses thermodynamic information and the centroid as a reference structure. It adds the prediction of the native structure into HelixPSO v1 and proposes a parallel version. IPSO introduces a specific objective as a function of MFE, number of selected stems and average length of selected stems. The approach is performed in three stages i.e. stage one is related to encode the source sequences by selecting all the legal stems and prepare a set of them. IPSO is employed in second stage for selecting the structures followed by obtaining the optimal secondary structure in stage three. FPSO incorporates two fuzzy logic controllers into PSO to adjust the inertia weight and learning factors. The aim is to balance between exploration and exploitation of PSO in the search space so as to improve solutions and avoid premature convergence. The study is based on searching optimal stem set. PSOfold is proposed by employing some modifications in IPSO related to enhancing the searching ability of optimal solution; settling the stem permutation problem. Two other competitive metaheuristics for RNA secondary structure prediction include GA based RNAPredict [18,19] and simulated annealing based SARNA-Predict [20]. RNAPredict finds RNA conformations at MFE by applying selection, mutation and crossover operators to population of chromosomes, whereas, SARNA-Predict employs the paradigm of cooling a substance that enhances free movement of particles at high temperatures.

Although a lot of computational approaches have been developed so far but an efficient computational approach is yet required to predict accurate RNA secondary structure. The problem of predicting secondary structure of RNA is very complex due to self-folding tendency and less stability of RNA, that enhances existence of several possible combinations of helices, bulges and loops. Hence, an approach is required that may decrease the problem complexity. On the other hand, the stability of the structure is related to the Gibbs free energy but not exactly obtained at the MFE. Hence, an approach is needed that performs more possible combinations of nucleotides for a structure close to MFE. Proposed work applies thermodynamic parameters based set-based approach. The use of different thermodynamic models plays significant role in improving the prediction accuracy, this idea has been verified in earlier works too [20,18]. RNAPredict [18] compares the improvement in the prediction accuracy with two thermodynamic models individual nearest neighbor (INN) and individual nearest neighbor-hydrogen bond (INN-HB). RNAPredict finds a significant correlation between lower free energy and increase in true positive base pairs. The correlation consecutively gets decreased as soon as the sequence length increases. Hence, it finds better prediction accuracy for suboptimal structures resulting to be more similar to the original structure than the optimal one. Similarly, SARNA-Predict [20] verifies the significance of thermodynamics models in accuracy improvement in the prediction of RNA secondary structure for which the secondary structure is encoded by permutations. The permutation has also been implemented in the proposed work using thermodynamic major model. Major model provides different combinations of helices that result in lower free energy suboptimal structures.

The approach in proposed work is employed for two levels i.e. in first level a specific objective function with hydrogen bond based parameters is optimized. Hence obtained *gbest* solutions are carried forward to second level for a nearest neighbor database (NNDB) parameters based MFE oriented objective function. Obtained results are compared with the family of PSO based algorithms, other popular metaheuristics and state-of-the-art secondary structure prediction algorithms.

The work is classified as follows: Section 2 delineates the basics of RNA secondary structure. Section 3 presents the introduction and algorithm of PSO followed by the details of TL-PSO algorithm applied for secondary structure prediction i.e. TL-PSOfold. Section 4 presents the experimental setup for benchmark dataset and algorithm parameters for TL-PSOfold. Section 5 discusses the results obtained, followed by the conclusions in Section 6.

2. RNA secondary structure prediction at minimum free energy

2.1. RNA secondary structure

RNA is a single-stranded chain of nucleotides adenine (A), uracil (U), guanine (G) and cytosine (C). The complimentary nucleotides are connected by hydrogen bonds. Watson-Crick base pairs are 'A' complimentary to 'U' (or vise versa), 'G' complimentary to 'C' (or vise versa) and Wobble base pair is 'G' complimentary to 'U' (or vise versa). These pairs are known as canonical pairs. A simple string of nucleotides A, U, G and C is known as the primary structure, whereas, after folding onto it, the complementary base pairs in RNA form a shape known as the secondary structure. RNA is different from DNA in the sense: DNA is double stranded, whereas RNA is single stranded; thymine in DNA is replaced by uracil in RNA; DNA contains deoxyribose sugar, whereas, RNA contains ribose sugar. Because of the existence of hydroxyl groups only in the ribose sugar, RNA is less stable than DNA, since hydroxyl groups are likely to hydrolyse [21]. RNA sequences have tendency to fold partially and pair with themselves to form double helices. This folding forms different secondary structures of RNA sequences. Secondary structure mainly consists of helices, bulges, internal loops, hairpin loops, dangling ends and multi-branch loops as depicted by Fig. 1. This structure is non-pseudoknotted structure of the RNA presented in [22]. For any two complementary base pairs $[a\ b]$ and $[c\ d]$ each attached by hydrogen bond, where a, b, c and d are the positions of nucleotide from 5′ end such that a is paired with b and c is paired with d, shown by $a < b$ and $c < d$.

A valid secondary structure should satisfy the following conditions:

1. A nucleotide a can pair only with one nucleotide (other than its neighbor at $a-1$ and $a+1$ position) i.e. $[a\ b] \bigcap [c\ d] = \phi$.
2. No pseudo-knots allowed i.e. if $a < c < b$, then $a < c < d < b$.

The basic structures formed in a folded RNA sequence along with respective notations as mentioned in Fig. 1 are:

- A structure formed by two strands that are attached by complementary base pairs is known as duplex.
- A base pair stacking is formed by two consecutive base pairs (a, b) and $(a+1, b-1)$.