Swarm and Evolutionary Computation xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Multi-objective two-level swarm intelligence approach for multiple RNA sequence-structure alignment

Soniya Lalwani[a], Rajesh Kumar[b,*], Kusum Deep[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, Indian Institute of Technology, Roorkee, India

ARTICLE INFO

Keywords:
Multi-objective optimization
RNA secondary structure
Multiple sequence alignment
Particle swarm optimization
Non-dominated solutions
Pareto optimal solution
Minimum free energy
Conflicting objectives

ABSTRACT

This paper proposes a novel two-level particle swarm optimization algorithm for multi-objective optimization (MO-TLPSO) employed to a challenging problem of bioinformatics i.e. RNA sequence-structure alignment. Level one of the proposed approach optimizes the dimension of each swarm which is sequence length for the addressed problem, whereas level two optimizes the particle positions and then evaluates both the conflicting objectives. The conflicting objectives of the addressed problem are obtaining optimal multiple sequence alignment as well as optimal secondary structure. Optimal secondary structure is obtained by TL-PSOfold, the structure is further used for computing the contribution of base pairing of individual sequence and the co-variation between aligned positions of sequences so as to make the structure closer to the natural one. The results are tested against the popular softwares for pairwise and multiple alignment at BRAlibase benchmark datasets. Proposed work is so far the first multi-objective optimization based approach for structural alignment of multiple RNA sequences without converting the problem into single objective. Also, it is the first swarm intelligence based approach that addresses sequence-structure alignment issue of RNA sequences. Simulation results are compared with the state-of-the-art and competitive approaches. MO-TLPSO is found well competent in producing pairwise as well as multiple sequence-structure alignment of RNA. The claim is supported by performing statistical significance testing using one way ANOVA followed by Bonferroni post-hoc analysis for both kind of alignments.

1. Introduction

Multi-objective optimization refers to the optimization of more than one objectives that are generally in conflict to each other, that results in no single solution for the problem. For such kind of optimization problems, the aim becomes to find good “trade-off” solutions that represent the best possible compromises among the objectives. Most of the real-world problems are multi-objective that belong to engineering, industries, biology, management and environment sciences. Recently, multi-objective optimization has shown significant impact on bioinformatics and computational biology. A few examples include inverse modeling, structure prediction and design, system optimization and experimental design, classification and sequence-structure alignment [1]. Proposed work belongs to the last category i.e. sequence-structure alignment that is performed here for ribose nucleic acid (RNA) with the help of swarm intelligence based technique.

RNA is a fundamental building block for all living cells. It plays numerous distinct and imperative roles in cell functioning i.e. protein synthesis, DNA replication, regulation and gene expression. The functions of RNAs are strongly related to their secondary structures due to the evolutionary conserved parts. Hence, alignment of RNA is essentially a multi-objective optimization problem (MOP), since contributions from both sequence similarity and secondary structure are needed to be taken into account. The NP-hardness of the problem along with conflicting objectives makes this job intricate and requires a suitable approach dealing with both the objectives efficiently.

Particle swarm optimization (PSO) is a stochastic metaheuristic, implemented to solve complex nature problems of this kind having features like: simple concept, easily implementable, robustness for parameter control and better computational efficiency than many other mathematical algorithms & heuristic optimization techniques [2]. Hence PSO is competently acquiring increasing attention and has become very popular approach for solving complex problems of MOP [3].

During literature survey, approaches found that are implemented for pair-wise and multiple RNA sequence-structure alignment include:

* Corresponding author.
E-mail addresses: slalwani.math@gmail.com (S. Lalwani), rkumar.ee@gmail.com (R. Kumar), kusumdeep@gmail.com (K. Deep).

http://dx.doi.org/10.1016/j.swevo.2017.02.002
Received 5 January 2016; Received in revised form 9 December 2016; Accepted 12 February 2017

Formal grammar based [4–7] that use stochastic context free grammars (SCFGs) for modeling alignment and folding processes; Linear integer programming [8] that proposes LARA (Lagrangian relaxation technique) by modeling the problem as an integer linear programming problem; Simulated annealing [9] that employs Markov chain Monte Carlo in a simulated annealing framework; Evolutionary approach i.e. Genetic algorithm based approaches Cofolga2 and Cofolga2mo [10,11]; Well established dynamic programming approach [12–16]; Graph representation technique [17] derived from RNA consensus abstract shapes; Stochastic sampling [18] based on searching the common structure between two sequences by probabilistically sampling (aligned stems based on stem conservation); Max-margin model [19] that automatically learns the weights for parameter estimation from training data; Maximum expected accuracy [20] estimator for maximization of the expected sum-of-pairs score of predicted alignment; Bayesian Markov chain Monte Carlo (MCMC) framework implementation [21] to sample from the joint posterior distribution of RNA triples. The details of all these RNA sequence-structure alignment techniques can be found in author's review paper [22].

Most recent work on RNA sequence-structure alignment includes as follows: ExpaRNA-P [23] computes exactly matching sequence-structure motifs in entire Boltzmann-distributed structure ensembles of two RNAs and implements ensemble-based sparsification so as to reduce problem complexity. It is based on the strategy of simultaneously match and fold RNAs. ReformAlign [24] presents a meta-alignment approach which has the procedure of using an existing alignment to construct a standard profile which summarizes the initial alignment, followed by realigning all sequences individually against the formed profile. Mattei et al. [25] presents a simple workflow for choosing suitable method for RNA pairwise alignment that depends on the input RNA primary sequence identity and the availability of reliable secondary structures. Performances of six different approach based algorithms is tested on datasets created by merging publicly available datasets of RNA secondary structure annotations with datasets of curated RNA alignments. HAlign [26] develops two MSA tools based on the centre star strategy. The first tool employs trie trees to accelerate the MSA of highly similar DNA sequences that is not for the big data. The second tool applies parallelism using the Hadoop platform to address large-scale datasets. Will et al. [27] presents SPARSE for simultaneous alignment and folding of RNAs by combining flexibility of the Sankoff model with lightweight computation. Bourgeade et al. [28] proposes a method namely RNA-unchained that addresses the general problem of the one-against-all RNA pairwise comparison, where a given query RNA is compared to unstructured set of target RNAs for computing high quality alignments. SARA-Coffee [29] is a tertiary structure-based multiple RNA aligner, at the concept of R-Coffee which is a consistency aligner for RNA. It is actually a modified version of T-Coffee with added predicted secondary structure feature. It is a suitable companion tool for any modeling technique that can benefit from a structurally accurate RNA MSA, including construction of profile SCFG. SARA-Coffee web server [30] is a part of the T-Coffee web platform, allows the online computation of 3D structure based multiple RNA sequence alignments. Lavender et al. [31] describe a dynamic programming approach for model-free sequence comparison that incorporates high-throughput chemical probing data. Based only on SHAPE [32] probing data, quite accurate alignment of ribosomal RNAs (rRNAs) from three diverse organisms (C. difficile, the archeon H. volcanii and the eubacteria E. coli) is performed. Consideration of base sequence identity and chemical probing reactivities improves accuracies further. Cech et al. [33] developed an algorithm for RNA pairwise structure superposition called SETTER, then extended it to the alignment of multiple RNA structures and developed the MultiSETTER algorithm. Algorithm web server provides both 3D graphics and summary statistics of the alignments. Swarm intelligence based RNA sequence-structure alignment technique, based on forming a context-sensitive hidden Markov model is developed in [6] which is a data training approach. To the best of authors' knowledge, [10,16] are the only multi-objective optimization (MOO) based approaches for RNA sequence-structure alignment, but they address only pair-wise alignment problem due to the complexity issues. Proposed approach is the first attempt for multiple sequence-structure alignment of RNA formulated as a multi-objective problem.

The optimal secondary structure for each sequence is obtained at minimum free energy (MFE) by TL-PSOfold [34] (two-level PSO algorithm for optimal folding of RNA), our previously developed algorithm. TL-PSOfold works in two levels: level one optimizes the number of stacked base pairs at hydrogen bond model parameters and level two minimizes free energy by following nearest neighbor database (NNDB) parameters. After determination of secondary structure of each sequence by TL-PSOfold, level one of proposed MO-TLPSO optimizes the length of gapped sequence. In level two, the optimal alignment with suitable gap positions is obtained at pre-determined sequence length by level one. For aligned sequences, secondary structure score is obtained by two parameters: base pair probability score of individual sequence and co-variation score between each pair of sequences. At the end of each iteration (in level two) two conflicting objectives remain to get optimized i.e. multiple sequence alignment (MSA) score and secondary structure (SS) score. The reasons of these two objectives to be conflicting is provided in Section 3. There is a trade-off between these two objectives so the Pareto optimality criteria with two level non-dominated sorting algorithm and crowding distance measure is employed here which stores the non-dominated and less crowded solutions in external archive.

The work is classified as follows: Section 2 presents the details of the objectives, sequence-structure alignment problem and two-level PSO (TL-PSO) algorithm. Section 3 carries the details of MO-TLPSO algorithm employed for sequence-structure alignment along with Pareto optimality criteria with certain features. Section 4 presents the details of experimental setup for algorithm and benchmark dataset. Section 5 discusses the results obtained, followed by the conclusions in Section 6.

2. Problem description and objectives

2.1. Particle Swarm Optimization

PSO is a derivative free class of population-based stochastic global optimization heuristic, firstly introduced by Kennedy and Eberhart [35] in 1995 for simulating social behaviour, as embodiment of the movement of organisms in a bird flock or fish school. These particles move in different directions in search of optimal solution, updating their positions and velocities through interaction for improving corresponding solutions. The objective function tending to find a parameter set $\vec{x}$ can be formulated as:

$$\min f(x) \quad \text{s. t.} \quad x \in S \subseteq R^D \tag{1}$$

where x is a matrix containing decision variables, composed of n vectors defined as $x=[\overrightarrow{x^1}, \overrightarrow{x^2}, \ldots \overrightarrow{x^n}]$ with dimension D. S is the feasible solution space of the problem. At t^{th} iteration, the previous velocity $v^i(t)$ and position $x^i(t)$ are updated as follows:

$$v^i(t+1) = wv^i(t) + c_1r_1(pbest^i(t) - x^i(t)) + c_2r_2(gbest(t) - x^i(t)) \tag{2}$$

$$x^i(t+1) = x^i(t) + v^i(t+1) \text{ with } x^i(0) \sim U(x_{\min}, x_{\max}) \tag{3}$$

Here, w is the inertia weight with constraint $0 \le w \le 1$ [36]. c_1 is the cognitive acceleration coefficient, whereas, c_2 is the social acceleration coefficient with constraint $c_1+c_2 \le 4$. r_1 and r_2 are uniform random numbers in range [0 1] [37]. $pbest^i$ is particles's own best performance defined by personal best i.e. $p_1^i, p_2^i, \ldots p_D^i$. Eq. (3) shows the position update $x^i(t)$ at iteration t for i^{th} particle, obtained by adding the updated velocity to its current position. $pbest$ at iteration $t+1$ is