Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Biology and Chemistry 29 (2005) 111–119

Computational Biology and Chemistry

www.elsevier.com/locate/compbiolchem

Cofolga: a genetic algorithm for finding the common folding of two RNAs

Akito Taneda *

Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan

Received 21 December 2004; accepted 22 February 2005

Abstract

In order to predict non-coding RNA genes and functions on the basis of genome sequences, accurate secondary structure prediction is useful. Although single-sequence folding programs such as mfold have been successful, it is of great importance to develop a novel approach for further improvement of the prediction performance. In the present paper, a secondary structure prediction method based on genetic algorithm, Cofolga, is proposed. The program developed performs folding and alignment of two homologous RNAs simultaneously. Cofolga was tested with a dataset composed of 13 tRNAs, seven 5S rRNAs, five RNase P RNAs, and five SRP RNAs; as a result, it turned out that the average prediction accuracies for the tRNAs, 5S rRNAs, RNase P RNAs, and SRP RNAs obtained by Cofolga with an optimal weight factor and default parameters were 83.6, 81.8, 73.5, and 67.7%, respectively. These results were superior to those obtained by a single-sequence folding based on free-energy minimization in which corresponding average prediction accuracies were 52.4, 47.4, 57.7, and 52.3%, respectively. Cofolga has a post-processing in which a single-sequence folding is performed after fixation of a predicted common structure; this post-processing enables Cofolga to predict a structure that is present in one of two RNAs alone. The executable files of Cofolga (for Windows/Unix/Mac) can be obtained by an e-mail request.

Keywords: Secondary structure prediction; Free-energy minimization; Sequence alignment; Covariation; Simulated annealing

1. Introduction

Non-coding RNAs (ncRNAs), which are the RNAs not translated to a protein, usually have their own characteristic secondary structures in accordance with their functions. Since the secondary structures play an important role in the analysis and prediction of the genes and functions of ncRNAs, various methods and programs for predicting the secondary structures have been developed. For example, free-energy minimization (FEM) (Zuker, 2003; Mathews et al., 1999) and covariation analysis (Chiu and Kolodziejczak, 1991; Gutell et al., 1992) have been widely used for the purpose.

Combination of different approaches is an attractive idea since it has a possibility to drastically improve the prediction accuracy. Sankoff algorithm is a dynamic programming approach for obtaining the simultaneous solution to the secondary structure prediction and the alignment of RNAs (Sankoff, 1985). In Sankoff algorithm, FEM and alignment (including covariation) are simultaneously taken into account by optimizing a hybrid objective function that is defined by a linear combination of an alignment-score term and a free-energy term. In the present paper, we call a secondary structure prediction which predicts the common secondary structure of a set of RNAs on the basis of FEM ‘common folding prediction’ (of course, including Sankoff algorithm). Since Sankoff algorithm is $O(N^4)$ in time and $O(N^6)$ in space, other common folding programs have been developed for more practical use: FOLDALIGN (Gorodkin et al., 1997) and RNAGA (Chen et al., 2000) are those for multiple sequences; Dynalign (Mathews and Turner, 2002) and CARNAC (Perriquet et al., 2003) are common folding programs developed for pairwise comparison. The common folding programs can be divided into two types in accordance with their purpose: Dynalign and RNAGA predict the whole structure of RNAs; FOLDALIGN and CARNAC are the pro-

* Tel.: +81 172 39 3662; fax: +81 172 39 3662.
E-mail address: taneda@si.hirosaki-u.ac.jp.

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.02.004

grams for finding the local structures such as hairpin loops. In addition to the methods mentioned above, alignment folding, e.g. alifold (Hofacker et al., 2002) and the method included in X2s package (Juan and Wilson, 1999), have been proposed for predicting the common structure of a set of RNAs with a pre-defined sequence alignment of the RNAs. Any common folding program has advantages and disadvantages. For example, while Dynalign is the algorithm closest to Sankoff's one and is a very accurate algorithm, it needs a large memory proportional to N^2, where N is a sequence length (e.g. Dynalign needs 256 MB to perform the common folding of sequences of 218 and 234 nt) (Mathews and Turner, 2002); although alignment folding is fast, it cannot improve the alignment when the given alignment is not accurate.

In the present paper, we propose a genetic algorithm (GA), common folding by genetic algorithm (Cofolga), for finding the common folding of two homologous RNAs. Cofolga predicts the whole common secondary structure of given RNAs by optimizing a function that has a hybrid form. In addition, by using a post-processing, Cofolga can also predict a non-common structure that exists in one of the RNAs alone. In the present study, pairwise common folding is carried out for tRNAs, 5S rRNAs, RNase P RNAs, and SRP RNAs to evaluate the performance of the proposed method; from the results of the tests, it is shown that the proposed algorithm improves the accuracy of the secondary structure prediction compared with a single-sequence folding method.

2. Algorithm

In Cofolga algorithm, the common folding prediction problem of two RNAs is solved by GA. Our GA is derived from RAGA (Notredame et al., 1997) that is based on the simple GA described in (Goldberg, 1987). GA searches for the solution with the highest objective function (OF) by iteratively updating a population of individuals (solutions) with various GA operators such as crossovers and mutations. In the present GA, alignments are used as an individual in the population and the size of the population is fixed through a run. The OF, f, used in Cofolga is as follows:

$$f = s + \frac{w}{N}\sum_{i=1}^{N} \Delta G_i, \qquad (1)$$

where s is an alignment score, w is a weight factor, N is the total number of RNA sequences ($N = 2$ in the present study), and ΔG_i is the free energy of the ith RNA. It is noted that the w must be negative, since lower free energy is more favorable. The schematic flowchart of Cofolga algorithm is shown in Fig. 1. As shown in Fig. 1, Cofolga comprises three GA steps (initialization, evaluation, reproduction) and a post-processing step. The detail of each step will be described in the next sections. The evaluation and reproduction steps repeats until the number of generations reaches the pre-defined maximum number (the population generated at nth GA step is called nth generation). As a maximum iteration number of the GA loops, 20 or 50 is used in accordance with the sequence length of the input sequences. A user can change the parameters of Cofolga such as the maximum iteration number and the population size.

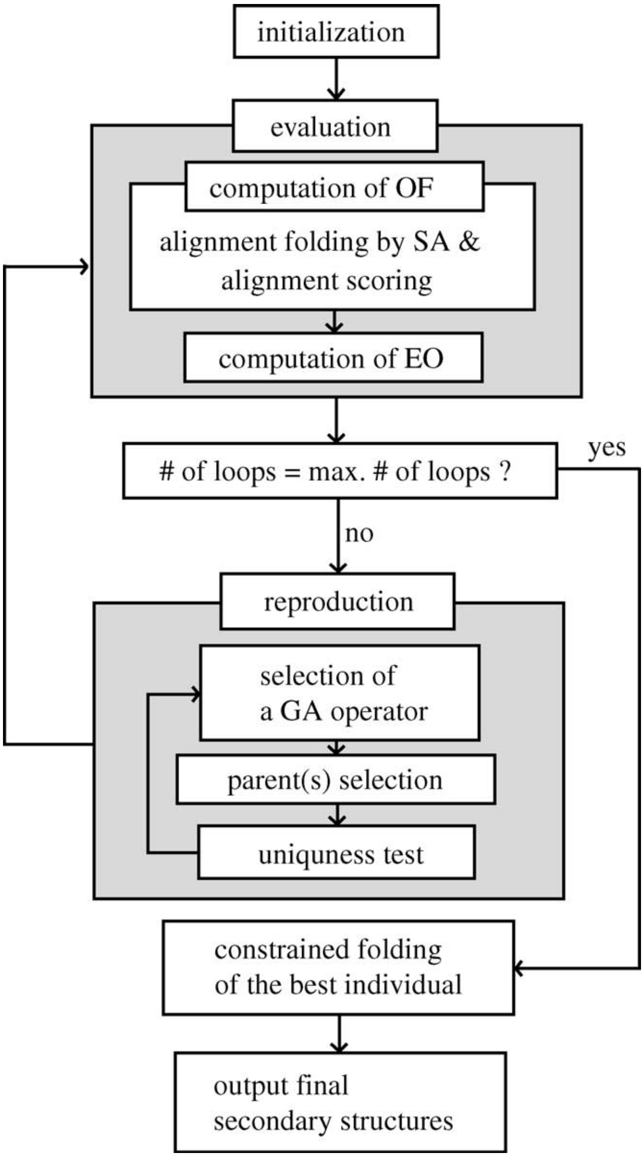

Fig. 1. A schematic flowchart of Cofolga algorithm.

2.1. Initialization

In initialization step, an initial population is generated by using dynamic programming with added noise (DPAN) (Gerstein and Levitt, 1996). First, individuals (alignments) are randomly generated by DPAN to fill the initial population. Then, an individual which is not unique in the population is removed. If there is a vacancy in the initial population after the procedure with DPAN, the following three-step procedure is invoked to fill the vacancy: (i) two sequences are randomly aligned without gap opening (terminal gap only);