Engineering Applications of Artificial Intelligence ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

A multiobjective swarm intelligence approach based on artificial bee colony for reliable DNA sequence design

José M. Chaves-González*, Miguel A. Vega-Rodríguez, José M. Granado-Criado

University of Extremadura, Department of Computers and Communications Technologies, Escuela Politécnica, 10003 Cáceres, Spain

ARTICLE INFO

Article history:
Received 21 July 2012
Received in revised form
8 January 2013
Accepted 22 April 2013

Keywords:
Multiobjective optimization
Swarm intelligence
Artificial bee colony
DNA sequence design
Multiobjective evolutionary algorithm
Molecular computing

ABSTRACT

The design of reliable DNA sequences is crucial in many engineering applications which depend on DNA-based technologies, such as nanotechnology or DNA computing. In these cases, two of the most important properties that must be controlled to obtain reliable sequences are self-assembly and self-complementary hybridization. These processes have to be restricted to avoid undesirable reactions, because in the specific case of DNA computing, undesirable reactions usually lead to incorrect computations. Therefore, it is important to design robust sets of sequences which provide efficient and reliable computations. The design of reliable DNA sequences involves heterogeneous and conflicting design criteria that do not fit traditional optimization methods. In this paper, DNA sequence design has been formulated as a multiobjective optimization problem and a novel multiobjective approach based on swarm intelligence has been proposed to solve it. Specifically, a multiobjective version of the Artificial Bee Colony metaheuristics (MO-ABC) is developed to tackle the problem. MO-ABC takes in consideration six different conflicting design criteria to generate reliable DNA sequences that can be used for bio-molecular computing. Moreover, in order to verify the effectiveness of the novel multiobjective proposal, formal comparisons with the well-known multiobjective standard NSGA-II (fast non-dominated sorting genetic algorithm) were performed. After a detailed study, results indicate that our artificial swarm intelligence approach obtains satisfactory reliable DNA sequences. Two multiobjective indicators were used in order to compare the developed algorithms: hypervolume and set coverage. Finally, other relevant works published in the literature were also studied to validate our results. To this respect the conclusion that can be drawn is that the novel approach proposed in this paper obtains very promising DNA sequences that significantly surpass other results previously published.

1. Introduction

Molecular computing, or more specifically DNA computing, refers to a computational model which uses DNA molecules as information storage and their biological reactions as processing operators (Adleman, 1994). In DNA computing, the hybridization between a specific DNA sequence and its base-pairing complement (also known as Watson–Crick pairing) is considered a crucial task, because this process makes possible to retrieve the information stored in DNA sequences and to perform operations with that information (Garzon and Deaton, 1999). However, undesirable hybridizations usually lead to incorrect computations. For this reason, DNA sequences generated for molecular computing have to be carefully designed. Thus, although DNA computing is a promising paradigm which is supposed to replace silicon-based computers in future decades, it presents some technical drawbacks that must be overcome. In fact, when these technical problems are solved, DNA computing will become an important area within Computer Engineering. Therefore, much work has focused on improving the reliability and efficiency of DNA computing, and particularly, on the design of error-minimized DNA sequences that are able to reduce the possibility of illegal reactions (Brenneman and Condon, 2002). The design of reliable DNA sequences which generate specific duplexes during hybridization, while simultaneously avoiding other undesirable cross-hybridization, involves several heterogeneous and conflicting design criteria which cannot be tackled by traditional optimization methods described in the related literature. Typical existing approaches for DNA sequence design problem include a wide range of non-exact approaches, such as evolutionary algorithms, dynamic programming, and heuristic methods (Brenneman and Condon, 2002). However, a design based on multi-objective evolutionary algorithms (MOEAs) represents the most appropriate design alternative (Coello et al., 2002; Deb, 2001) because MOEAs take into account several conflicting objectives simultaneously without

* Corresponding author. Tel.: +34 678962634.
E-mail addresses: jm@unex.es, jmanuel.gonzalez@gmail.com (J.M. Chaves-González), mavega@unex.es (M.A. Vega-Rodríguez), granado@unex.es (J.M. Granado-Criado).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.04.011

the artificial adjustments which are included in classical mono-objective optimization methods. These approaches usually include the use of different weighting factors assigned to each objective.

This paper proposes a new DNA sequence design evolutionary approach based on multiobjective swarm intelligence to automatically generate reliable DNA strands that can be applied to molecular computing (a promising area in Computer Engineering). Specifically, the artificial bee colony algorithm (Karaboga and Akay, 2009) was adapted with the inclusion of several multiobjective features, such as the idea of non-dominated solutions or the concept of non-dominated sorting (Deb et al., 2001). Furthermore, our approach is compared against the well-known multiobjective standard NSGA-II (fast non-dominated sorting genetic algorithm (Deb et al., 2001)). Finally, our results are validated by using other works published in the literature. As will be discussed, our approach generates more reliable DNA sequences for DNA computing than other relevant approaches previously published.

The rest of this work is organized as follows: Section 2 discusses related work. Section 3 describes the basic background on the problem and the multiobjective formulation followed. The metaheuristics which have been developed are explained in Section 4. Section 5 is devoted to present and analyze the results. In Section 6, our approach is compared with other methods published in the literature. Finally, Section 7 summarizes the conclusions of the paper.

2. Related work

DNA computing has been a very active research field in the last decade. There are a great number of works in which different approaches have been proposed to generate reliable DNA sequences that are suitable for bio-molecular computing. However, most of related publications manage the problem in terms of threshold-based constraints by considering different biological criteria as requirements which are combined into a final mono-objective function. Table 1 summarizes a review of recent DNA sequence generators classified according to the kind of methodology used to generate sequences.

Exhaustive and random searches (Hartemink et al., 1998; Penchovsky and Ackermann, 2003) represent the simplest methods, but they are not effective because they use a great amount of computational resources. Template-map strategies (Frutos et al., 1997; Arita and Kobayashi, 2002; Liu et al., 2003) are designed to choose dissimilar sequences among a huge set of sequences automatically generated. In Feldkamp et al. (2001), the design of a limited number of sequences is performed by using a representation based on directed graphs. In this approach, graph nodes represent base strands in which each node has four strands that can appear as successors in a longer sequence. Tanaka et al. (2001) generated sequences by using simulated annealing. They tried to find promising solutions by combining different criteria into a fitness function. Dynamic programming was proposed in Marathe et al. (1999) to design DNA sequences based on Hamming distance and free energy. Contrary to the system referred previously, biologically inspired and evolutionary methods have been recently used. Deaton et al. (2002) proposed a method based on in-vitro evolution to find non-cross-hybridizing DNA libraries. However, biologically inspired methods have some inherent problems, such as they are not able to distinguish between DNA sequences in the library. Other biological approaches consider thermodynamic properties presented in DNA structures or free energy of DNA sequences (Deaton et al., 2002; Heitsch et al., 2002).

However, according to the vast number of works published in the last years (Deaton et al., 1998; Ibrahim et al., 2012), evolutionary algorithms (EAs) can be considered the most widely applied techniques for designing reliable DNA sequences. EAs use one or more design criteria to make evolve the particular evolutionary scheme adopted. For example, genetic algorithms (GAs) have been frequently used due to their simplicity (Deaton et al., 1998; Ruben et al., 2001). In Deaton et al. (1998), the fitness function was designed by using the Hamming distance between sequences. An iterative genetic search was developed in (Zhang and Shin, 1998) to design sequences in the context of DNA computing. Several fitness criteria (similarity, H-measure, hamming distance and GC ratio) were applied in a constrained genetic algorithm in Arita et al. (2000). Shin et al. (2002, 2005) proposed one of the few multiobjective evolutionary algorithms published in the literature. For this reason, our work is compared against these studies in Section 6. A multiobjective approach based on six DNA design criteria was proposed in Shin et al. (2002) and improved in Shin et al. (2005). Two criteria (melting temperature and GC ratio) were considered as constraints meanwhile similarity, H-measure, continuity and hairpin were the four objectives of the system. Other studies also consider several design criteria, but they eventually manage them into a weighted mono-objective function. Thus, in Xu et al. (2008) and Cui and Li (2010), DNA sequence design problem is formulated as a multi-criteria optimization problem tackled with Genetic Algorithm (GA)/Particle Swarm Optimization (PSO). In Khalid et al. (2008), PSO is applied

Table 1
DNA sequence design approaches found in the literature.

Proposed methodology	References and year of publication
Exhaustive search	Hartemink et al. (1998)
Random search	Penchovsky and Ackermann (2003)
Template map strategies	Frutos et al. (1997), Arita and Kobayashi (2002) and Liu et al. (2003)
Graph method	Feldkamp et al. (2001)
Stochastic strategy	Tanaka et al. (2001)
Dynamic programming	Marathe et al. (1999)
Biologically inspired algorithms	Deaton et al. (2002a, 2002b) and Heitsch et al. (2002)
Evolutionary algorithms (EAs)	Deaton et al. (1998) and Zhang and Shin (1998) Arita et al. (2000) Ruben et al. (2001) Shin et al. (2002) Shin et al. (2005) Xu et al. (2008), Khalid et al. (2008) and Kurniawan et al. (2008, 2009) Hongyan and Xiyu (2009) and Wang et al. (2009) Zhang et al. (2010) and Cui and Li (2010) Xiao and Cheng (2011) and Mustaza et al. (2011) Ibrahim et al. (2012)

Please cite this article as: Chaves-González, J.M., et al., (2013), http://dx.doi.org/10.1016/j.engappai.2013.04.011