Expert Systems With Applications 56 (2016) 209–226

Contents lists available at ScienceDirect

Expert Systems With Applications

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

ADEMO/D: An adaptive differential evolution for protein structure prediction problem

Sandra M. Venske[a,*], Richard A. Gonçalves[a], Elaine M. Benelli[b], Myriam R. Delgado[c]

[a] Department of Computer Science, UNICENTRO, Guarapuava, Brazil
[b] Department of Biochemistry and Molecular Biology, UFPR, Curitiba, Brazil
[c] CPGEI/DAINF, UTFPR, Curitiba, Brazil

ARTICLE INFO

Article history:
Received 9 April 2015
Revised 14 February 2016
Accepted 4 March 2016
Available online 10 March 2016

Keywords:
Multiplicative optimization
Adaptive differential evolution
Probability matching
Protein structure prediction

ABSTRACT

Protein Structure Prediction (PSP) is the process of determining three-dimensional structures of proteins based on their sequence of amino acids. PSP is of great importance to medicine and biotechnology, e.g., to novel enzymes and drugs design, and one of the most challenging problems in bioinformatics and theoretical chemistry. This paper models PSP as a multi-objective optimization problem and adopts ADEMO/D (Adaptive Differential Evolution for Multi-objective Problems based on Decomposition) on its optimizer platform. ADEMO/D has been previously applied to multi-objective optimization with a lot of success. It incorporates concepts of problem decomposition and mechanisms of mutation strategies adaptation. Decomposition-based multi-objective optimization tends to be more efficient than other techniques in complex problems. Adaptation is particularly important in bioinformatics because it can release practitioners, with a great expertise focused on the application, from tuning optimization algorithm's parameters. ADEMO/D for PSP needs a decision maker and this work tests four different methods. Experiments consider off-lattice models and *ab initio* approaches for six real proteins. Results point ADEMO/D as a competitive approach for total energy and conformation similarity metrics. This work contributes to different areas ranging from evolutionary multi-objective optimization to bioinformatics as it extends the application universe of adaptive problem decomposition-based algorithms, which despite the success in various areas are practically unexplored in the PSP context.

1. Introduction

Proteins realize multi-objective most of vital structural, enzymatic, transport, and regulatory functions in the cell. Protein functions are determined by their structures, which can be organized into four levels of hierarchies with an increasing complexity: primary, secondary, tertiary, and quaternary structure (Bujnicki, 2009). The protein's native conformation (tertiary or quaternary) is often necessary to understand its function at a molecular level (Cohen & Kelly, 2003). Indeed, this is an important information for designing new enzymes or drugs for specific diseases (Hassanien, Milanova, Smolinski, & Abraham, 2008). Protein structure prediction (PSP) is an important research topic of Bioinformatics. The classical experiment performed by Christian Anfinsen in the early 1970s demonstrated the foundation of PSP: all necessary information to proteins fold properly is encoded in their primary structure (the amino acid or residue sequence) (Zaki & Bystroff, 2008).

A protein in a folded state achieves the lowest free energy conformation. Therefore, it seems natural to address the PSP problem first computing the free energy of every possible conformation and then selecting the structure corresponding to the lowest value. Finding the problem's solution encompasses two difficult tasks: searching candidate conformations and evaluating the free energy of a particular conformation.

In the first case, exhaustive search strategies are prohibitive because the number of conformations grows exponentially with the number of residues – several authors have proven that the PSP is an NP-hard problem (Fraenkel, 1993; Unger & Moult, 1993). In this case one must, instead, perform an approximate (stochastic) search where the better is the (sampling) of lower energy conformations, the better is the search method (Tramontano, 2006).

In the second case, one may also approximate the free energy calculation of a protein conformation. Usually, the approximation function is based on two energies values: bond and non-bond

* Corresponding author. Tel.: +55 4236298344.
E-mail addresses: ssvenske@unicentro.br, sandravenske@gmail.com (S.M. Venske), richard@unicentro.br (R.A. Gonçalves), benelli@ufpr.br (E.M. Benelli), myriamdelg@utfpr.edu.br (M.R. Delgado).

http://dx.doi.org/10.1016/j.eswa.2016.03.009
0957-4174/

atoms interactions. Some recent experimental researches indicate that these interactions are in conflict (Becerra, Sandoval, Restrepo-Montoya, & Ni no, 2010; Cutello, Narzisi, & Nicosia, 2006; Handl, Lovell, & Knowles, 2008), justifying a multi-objective formulation for the PSP problem.

In spite of the great number of researches, the PSP solution is still an open field (de Andrades, Dorn, Farenzena, & Lamb, 2013; Ding et al., 2012; Dorn, Buriol, & Lamb, 2013). Three-dimensional structures of some proteins can be determined experimentally, commonly, using X-ray crystallography and nuclear magnetic resonance spectroscopy. Although very precise, these methods are costly and may present some experimental limitations, such as the efficient and rational production of proteins with structural properties including high-throughput cloning and expression from multiple vectors in multiple host organisms, core domain identification using proteolysis methods, and the use of expression and detection tags (Liu & Hsu, 2005). Therefore, computational strategies have been developed to offer alternatives to solve the PSP problem (Tramontano, 2006). Evolutionary methods, specially evolutionary multi-objective algorithms in the last few years, appear as candidates. Differential Evolution (DE) (Storn & Price, 1997), including its multi-objective versions, are examples of these promising evolutionary methods.

PSP was first handled as a multi-objective optimization problem in (Angeline, Michalewicz, Schoenauer, Yao, & Zalzala, 1999) and it was solved by the evolutionary algorithm called Pareto Archived Evolution Strategy (PAES). In (Venske, Gonçalves, & Delgado, 2012a, 2014) we also proposed an evolutionary multi-objective algorithm, which served as the basis for the approach considered in this work. The approach combines ADE (Adaptive Differential Evolution) with the MOEA/D (Multi-objective Evolutionary Algorithm based on Decomposition) framework (Zhang & Li, 2007), which is well suited to solve complex problems (Li & Zhang, 2009; Lin et al., 2016) and therefore should be appropriate to solve PSP. Named ADEMO/D, it achieved very promising results for the addressed benchmark (indeed it outperformed the winner of CEC-2009 competition for most of considered instances). In (Venske, Gonçalves, & Delgado, 2014) we tested two methods to perform adaptive strategy selection: probability matching and adaptive pursuit. These selection methods were combined with four different rewarding techniques: average absolute, average normalized, extreme absolute and extreme normalized. We observed that the version using Probability Matching (PM) combined with Extreme Absolute (ExtAbs) reward obtained the best results. Therefore, these operators were chosen to compose the standard version of the ADEMO/D algorithm used in this work.

In this paper we extend the works presented in (Venske, Gonçalves, & Delgado, 2012a, 2012b; Venske, Gonçalves, Martin, & Delgado, 2013) and (Venske et al., 2014). As in (Venske et al., 2013), PSP assumes an off-lattice model where a protein is represented as a chain of residues or groups of residues moving through a continuous space. We also consider the challenging *ab initio* approach for PSP, that is a template-free modeling which is by now recognized as one of the most difficult problems in computational structural biology (Unger & Moult, 1993). In this paper we include four real proteins in addition to those considered in (Venske et al., 2013). Another contribution is the formalization of PSP as a continuous multi-objective optimization problem. Different from Venske et al. (2012b), here we use the problem decomposition platform (MOEA/D) as the optimizer framework. Aiming to adapt the approach proposed in (Venske et al., 2012a) and Venske et al. (2014) to PSP, a decision maker must be used to define which solution (protein conformation) is provided as the predicted structure, i.e., the decision maker is responsible for choosing the best solution (closest to the natural conformation), among all non-dominated solutions. Returning the best solution of a Pareto Front is a challenge for multi-objective optimization. While Venske et al. (2013) considered only one decision maker, this paper tests four different methods to define the final conformation.

Despite the success of adaptive evolutionary techniques (Dragoi & Dafinescu, 2016; Xu & Zhang, 2013), particularly in the DE literature (Das, Mullick, & Suganthan, 2016), and the fact that it alleviates the practitioner's task of finding good parameter settings, there are few works using adaptation for PSP. Some of them also use off-lattice models (Liu & Tao, 2006; Nicosia & Stracquadanio, 2007; Sudha, Baskar, & Krishnaswamy, 2013; Tantar, Melab, & Talbi, 2008). Different search strategies are used: simulated annealing algorithm (Liu & Tao, 2006; Tantar et al., 2008), generalized pattern search (Nicosia & Stracquadanio, 2007), mesh adaptive direct search (Nicosia & Stracquadanio, 2007) and differential evolution (Sudha et al., 2013). These works are briefly discussed in Section 2.3.

In summary, ADEMO/D has been chosen because it was successfully applied in (Venske et al., 2012a, 2014) on a well known set of multi-objective benchmarks (CEC-2009), and now we aim to extend its application universe. Besides previous favorable results, ADEMO/D is a multi-objective adaptive and decomposition-based approach which explores two currently promising DE research topics: parameter adaptation and multi-objective optimization (Das et al., 2016). So, this paper intends to investigate if PSP can benefit from the ADEMO/D's capacity of decomposing the problem into a set of subproblems and adjusting the search to different directions, based on the knowledge acquired during the evolutionary process.

The main contributions of this paper encompass different fields. It contributes (i) to adaptive differential evolution research as it evaluates the use of adaptation to find out solutions for the PSP problem; (ii) to multi-objective optimization as four decision makers are tested with ADEMO/D for PSP: two and four neighbors angle-based methods (Branke, Deb, Dierolf, & Osswald, 2004), a decision maker based on the total energy of protein and finally, one based on empirical point (Coello Coello, Lamont, & Van Veldhuizen, 2007). This is particularly important in the PSP context as, to the best of our knowledge, for the first time different decision makers are investigated while solving the multi-objective PSP formulation; (iii) to the universe application of decomposition-based algorithms as we extend their use to the PSP problem. This universe is practically unexplored - the resumed version of this paper (Venske et al., 2013) is the first work that uses problem decomposition to PSP; (iv) to the bioinformatic community as we are testing a number of different proteins and presenting a comparative result with the literature in terms of final potential energy and a classic metric of distance between protein conformations.

The remainder of this paper is organized as follows. Section 2 presents the PSP problem. In Section 3, we formalize PSP as a multi-objective optimization problem. Section 4 presents an overview of the fundamental concepts related to adaptive differential evolution. The proposed approach is detailed in Section 5. Experiments and results are presented and discussed in Section 6. Section 7 concludes the paper and discusses directions for the future work.

2. Protein structure prediction

The base of a protein is its *primary structure* (formed by the amino acids chains) which determines its different chemical properties. Amino acids rearrange themselves creating local folds, mostly α-helices and β-strands, called *secondary structure* (2D) of the polypeptide chain. A *tertiary structure* (native conformation for mono chain protein) is the arrangement of secondary structure elements in three dimensional spaces. Proteins assume a three-dimensional shape which usually determines their function