Computers and Mathematics with Applications 64 (2012) 3792–3804

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

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

A tree-structured covalent-bond-driven molecular memetic algorithm for optimization of ring-deficient molecules

M.M. Ellabaan [a,1], S.D. Handoko [a,1], Y.S. Ong [a,*], C.K. Kwoh [a], S.A. Bahnassy [b], F.M. Elassawy [c], H.Y. Man [d]

[a] *Centre of Computational Intelligence, School of Computer Engineering, Nanyang Technological University, 639798 Singapore, Singapore*

[b] *Biochemistry Department, Faculty of Science, Alexandria University, Alexandria, Egypt*

[c] *Central Lab, Faculty of Pharmacy, Alexandria University, Alexandria, Egypt*

[d] *Department of Biology, Boston University, Boston, MA, USA*

ARTICLE INFO

Keywords:
Evolutionary optimization
Molecular optimization
Memetic algorithms
Neuroscience
Glutamic acid
Stereoisomers

ABSTRACT

With enormous success in both science and engineering, the recent advances in evolutionary computation—particularly memetic computing—is gaining increasing attention in the molecular optimization community. In this paper, our interest is to introduce a memetic computational methodology for the discovery of low-energy stable conformations—also known as the stereoisomers—of covalently-bonded molecules, due to the abundance of such molecules in nature and their importance in biology and chemistry. To an optimization algorithm, maintaining the same set of bonds over the course of searching for the stereoisomers is a great challenge. Avoiding the steric effect, *i.e.* preventing atoms from overlapping or getting too close to each other, is another challenge of molecular optimization. Addressing these challenges, three novel nature-inspired tree-based evolutionary operators are first introduced in this paper. A tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA)—tailored specifically to deal with molecules that involve covalent bonding but contain no cyclic structures using the three novel evolutionary operators—is then proposed for the efficient search of the stereoisomers of ring-deficient covalently-bonded molecules. Through empirical study using the glutamic acid as a sample molecule of interest, it is witnessed that the proposed TCM-MA discovered as many as up to sixteen times more stereoisomers within as little as up to a five times tighter computational budget compared to two other state-of-the-art algorithms.

1. Introduction

Amino acids, which are the building blocks of proteins, are covalently bonded and are essential for life, constituting the structure and machinery of living organisms [1]. Most drugs – if not all – are also covalently bonded [2,3]. Recent studies show that different functions may be assumed as the molecules adopt various low-energy stable conformations, allowing selective or preferential interactions with different systems [4–7]. One particular example of such molecules is the glutamic acid, a major neurotransmitter in the central nervous system that plays a key role in brain functions including learning and memory and neurological disorders [8]. The glutamic acid can adopt several low-energy stable conformations, enabling it to selectively interact with glutamate receptors and transporters [9,10]. Some conformations are reported to activate glutamate receptors. Others are speculated to inhibit glutamate transporters. Over-activation or inhibition of the

* Corresponding author.
E-mail address: asysong@ntu.edu.sg (Y.S. Ong).

[1] These authors contributed equally to this work.

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.02.055

glutamate receptors or transporters, respectively, may escalate into neurotoxicity [11,12]. Drugs of the glutamate analogues, for instance, are more effective in some conformations than others. This suggests that identification of the stereoisomers, the various structural configurations of low-energy stable conformations that share the same set of covalent bonds, of the glutamic acid may help scientists in designing more effective drugs. At the same time, it may also help neuroscientists to both understand the functions of this neurotransmitter at the atomic level and better clarify its impacts on the functions and dysfunctions of the neurons.

Identifying stereoisomeric conformations thus advocates significant importance for revealing the structure–function relationship of biomolecules but poses a huge challenge to researchers. Wet-lab experiments are possible but they are normally expensive and require specially-tailored spectroscopic technology available at only a few finger-countable laboratories worldwide. The computational approach then provides the more affordable alternative. Covalent bonds, however, cause some difficulties for canonical optimization algorithms to maintain the same set of bonds over the course of searching for stereoisomers. Other difficulties include avoiding the steric effect, i.e. preventing atoms from overlapping or getting too close to each other. Expensive computation of the potential energy function, furthermore, restricts the number of evaluations possible over the course of searching for stereoisomers in order to keep the overall computation time reasonable. With a single evaluation lasting up to hours, the design of some efficient memetic computing methodology is therefore undoubtedly necessary. Memetic computation represents the most recent advances in evolutionary computation with enormous success in both science and engineering. Despite the success that the field has enjoyed, the design of specialized memetic methodology for molecular optimization of covalently-bonded molecules has not been commonly observed. In contrast to earlier works, introduced in this paper for the first time is the tree representation of a covalently-bonded molecule where connectivity information about the covalent bonding in the molecule is embedded; using which three novel nature-inspired tree-based evolutionary operators are then designed. A tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA) that is tailored specifically to deal with molecules that involve covalent bonding but contains no cyclic structures using the novel evolutionary operators is then proposed. This, in turn, allows the efficient search of stereoisomers of some ring-deficient covalently-bonded molecules. In particular, the proposed algorithm, in its course of evolution, produces mostly the stereoisomer candidates of the molecule and eliminates as many as possible of the nonsensical structures of the molecule of interest. As such, most efforts are thereby concentrated on exploring the feasible or near feasible search space. Empirical study using the glutamic acid [13] as a sample molecule of interest provides evidence that the proposed TCM-MA is indeed more efficient, discovering as many as up to sixteen times more stereoisomers within as little as up to a five times tighter computational budget, compared to two other competing algorithms.

In what follows, a formal formulation of the problem of finding stereoisomers of the covalently-bonded molecules shall first be presented in Section 2. The canonical memetic framework along with their evolutionary operators, existing niching strategies, and commonly-used individual-learning procedures are then presented in Section 3. Addressing the limitations of the conventional evolutionary operators in Section 3, three novel nature-inspired tree-based operators are then proposed in Section 4. Consolidating the three novel operators with the valley-adaptive clearing method as the niching strategy and the GEDIIS method of Gaussian 09 as the individual-learning procedure, the TCM-MA shall conclude Section 4. In Section 5, results from the empirical study of the proposed algorithm using the glutamic acid as a sample molecule of interest are then presented and discussed. Lastly, Section 6 concludes the contributions of the works presented in this paper and provides plausible future research directions.

2. Problem formulation

The energy landscape has proven itself as a useful underlying conceptual framework in fields like protein folding and docking as well as small molecule optimization [14,15]. It is formally defined as $\mathcal{L} = (\mathbf{X}, f, d)$ in which $\mathbf{X}$ is the set of all possible structural configurations of the molecule of interest, $f : \mathbf{X} \rightarrow \Re$ the potential energy function of any single structural configuration in $\mathbf{X}$, and $d : (\mathbf{X}, \mathbf{X}) \rightarrow \Re$ the distance measure between any two structural configurations in $\mathbf{X}$. Every configuration $\mathbf{x} \in \mathbf{X}$ of some molecule of interest is a vector of $3n$ real-valued variables representing the three-dimensional Cartesian coordinates – measured in Angstroms (Å) – of the n atoms that make up the molecule. It should be noted that $X \subset \Re^{3n}$ as not all vectors of $3n$ real values constitute the set of possible structural configurations of the molecule.

The isomers are defined as configurations that share the same set of atoms. Depending on whether or not these configurations also share the same set of bonds, they are classified into stereoisomers and constitutional isomers. Stereoisomeric configurations share the same set of bonds while constitutional isomers involve the process of breaking some bonds and the formations thereof. On the energy landscape $\mathcal{L}$, any stereoisomer $\mathbf{x}^*$ is defined mathematically as a stationary point where $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive-definite, hence one of the multiple minimum energy conformations ($\mathbf{x}^* \in \mathbf{X}^*$). This paper focuses on the discovery of as many as possible the stereoisomers of glutamic acid as the sample molecule of interest. Oftentimes, researchers are only interested in good stereoisomers that are characterized by the following properties [16].

- $\|\nabla f(\mathbf{x}^*)\| < \lambda$ where λ is an infinitesimally small precision tolerance.
- $\forall \mathbf{x} |f(\mathbf{x}) - f(\mathbf{x}^*)| < \varepsilon \Rightarrow d(\mathbf{x}, \mathbf{x}^*) < \gamma$ where $x \in \mathbf{X}^* \setminus \{\mathbf{x}^*\}$, ε and γ are the maximum acceptable similarities in the potential energy and the structural configuration space, respectively, and $d(\mathbf{x}, \mathbf{x}^*)$ is the USR structure dissimilarity [17–19] between structure $\mathbf{x}$ and $\mathbf{x}^*$.
- $f(\mathbf{x}^*) < f_{max}$ where f_{max} is a user-defined threshold of potential energy.