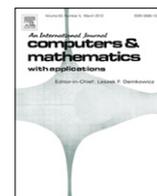


Contents lists available at [SciVerse ScienceDirect](http://SciVerse.Sciencedirect.com)

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. Ellassawy^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.

© 2012 Elsevier Ltd. All rights reserved.

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.

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, $f : \mathbf{X} \rightarrow \mathfrak{R}$ the potential energy function of any single structural configuration in \mathbf{X} , and $d : (\mathbf{X}, \mathbf{X}) \rightarrow \mathfrak{R}$ 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 $\mathbf{X} \subset \mathfrak{R}^{3n}$ 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 $\mathbf{x} \in \mathbf{X}^* \setminus \{\mathbf{x}^*\}$, 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.

3. Literature review

Memetic computing represents an emerging field within the context of evolutionary computation that has attracted ever increasing research attention in the past decades, supported by a growing number of reported successes [20]. Memetic computing was first introduced as the simple memetic algorithm (MA), the hybridization of population-level global search and individual-level local refinement [21]. Inspired by Darwin's theory of biological evolution and Dawkin's notion of cultural evolution, MAs facilitate two central yet competing goals of optimization, the search space exploration and exploitation, thus allowing rapid convergence to the precise local optima. To-date, MAs have frequently been applied to solving real-world problems more efficiently, particularly in various science and engineering design problems [22], including those with a medical nature. Employing MAs, higher-quality solutions are attained much faster than using traditional evolutionary algorithms [23–28].

3.1. Initialization

Several samples of the entire search space are produced and assigned as the initial population during the initialization. Therefore, a good sampling strategy is undoubtedly crucial. It is intuitive that too much bias towards a particular optimum in the initial population would cost the algorithm longer evolution time or even failure to locate the other optima. Due to the curse of dimensionality [29], a random initialization procedure based on uniformly distributed random numbers is generally favored. Using such a procedure, a random number is generated for each of the chromosomes of every individual in the initial population. When applied to covalently-bonded molecular systems, however, it is intuitive that such a random initialization procedure would easily produce nonsensical initial molecular structures as if covalent bonds are broken at places where there should be such bonds and formed at places where there should not be any of them.

3.2. Canonical evolutionary operators: Crossover and mutation

In canonical evolutionary algorithms, random changes are induced through two operators: crossover and mutation. On one hand, the crossover operator combines two parent chromosomes, based on the idea that the exchange of good genes between parents will eventually produce better offspring. Many crossover operators have been proposed in the last few decades. Among the most efficient are PBX [30], Laplace [31], and hill-climbing [32] crossovers. The mutation operator, on the other hand, alters slightly the value of one or more chromosomes of the individuals in the population in an analogous manner to the biological mutation. The mutation operator allows individuals to escape from some local optima to converge to other local optima or to the global optimum. Among the most efficient mutation operators proposed recently are power mutation [31], the Makinen–Periaux–Toivanen [33] mutation, and the uniform mutation [34]. However, these crossover and mutation operators interpret the chromosomes as some linear vectors, and therefore, may not be suitable for the optimization of the covalently bonded molecular system. Atoms that are supposedly close together due to some covalent bonds could be far apart after the recombination as if the bonds are broken. This creates unnecessary exploration of the search space as such structures have violated the covalent bonding of the original molecular structure of interest, and therefore, need not be considered in the optimization process.

3.3. Explorative and exploitative memetic operators: Niching and individual learning

Unlike conventional evolutionary computing paradigm, multimodal memetic paradigm provides two main operators, namely, niching and individual learning, to provide efficient exploration and exploitation of the landscape. Niching provides a means of finding and preserving multiple stable niches, or the favorable search space regions, so as to avoid convergence to a single solution. Among the most efficient and widely-used niching operators are the clearing [35], modified clearing [36] and the valley-adaptive clearing (VAC) [37] techniques. As finding multiple stereoisomers is of utmost interest, the VAC as the latest development of the clearing technique will be used throughout this paper. This is to complement the strengths of the proposed tree-based initialization, crossover, and mutation operators by maintaining the population diversity at each generation so as to find a larger number of stereoisomers within a smaller computational budget.

Additionally, the life-time learning of all or some individuals in the current population facilitates the goal of the search space exploitation of every optimization algorithm. Searching for an optimum within some locality, hence the alternative term “local search” or “local refinement”, a learning operator refines an individual and allows rapid convergence of that individual to some local optimum with sufficiently high precision. Individual learning operators include the first- and second-order methods. Utilizing the first-order derivatives information of the objective function is the widely-used steepest descent [38] method. Utilizing the additional second-order derivatives information of the objective function are the Newton–Raphson [39] and the rational function optimization [40] method. In the absence of the analytical gradient and Hessian matrix, the eigenvector-following optimization [41] method then estimates those quantities using a second-order Taylor expansion of the objective function. In molecular optimization, the first three methods are frequently used since the derivatives of the energy function are often available as they relate to quantities that are of interest to researchers in the field. Recently, a hybrid geometry optimization method with energy-represented direct inversion in the iterative subspace (GEDIIS) [42] was proposed. It minimizes an energy representation of the local potential energy surface in the vicinity of

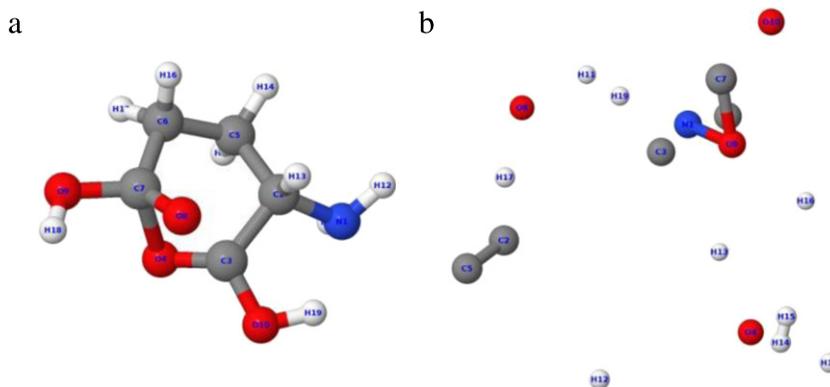


Fig. 1. Examples of non-stereoisomer candidates: (a) constitutional isomer candidate of glutamic acid; (b) nonsensical structure of glutamic acid.

previously evaluated structures as a least-square problem using some least-square minimization scheme and interpolation as well as direct inversion in the iterative subspace. Designed specifically for large, floppy molecules and readily available in the Gaussian 09 software package, the GEDIIS will be employed as the individual-learning operator throughout this paper.

4. Proposed methodology

The novelty of the proposed molecular memetic algorithm in this paper lies in the nature-inspired tree-based evolutionary operators, which are capable of exploring the structural configuration space of the ring-deficient covalently-bonded molecules more efficiently than any other canonical evolutionary operators. This is achieved by producing mainly realistic structures of the molecules with a large portion containing its stereoisomer candidates and a small percentage its constitutional isomer candidates. In the next section, this shall be witnessed through an empirical study of the proposed algorithm. Working together with valley-adaptive clearing method as the niching strategy and the GEDIIS method of Gaussian 09 as the individual-learning technique, the newly proposed tree-based evolutionary operators effectuate the tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA); the key components of which shall be elaborated in greater detail and discussed in the rest of this section.

4.1. Tree-based evolutionary operators

To recognize a molecule is covalently bonded constitutes the complete knowledge of all the covalent bonds present therein. Optimizing the structure of such a molecule in the search for its stereoisomers is greatly challenged by the maintenance of the existing set of covalent bonds without incurring any additional set of such bonds. In other words, the set of covalent bonds of the molecule must remain the same. Unfortunately, conventional evolutionary operators discussed in the previous section perform initialization as well as crossover and mutation without any awareness of being constrained by such bonding. As illustrated by the examples in Fig. 1, the resulting individuals can therefore be the nonsensical structures or the constitutional isomer candidates of the molecule. Impossible bond lengths, improbable bond angles, and/or steric clashes among atoms could have occurred very easily in the course of the evolution of the classical memetic algorithms. Involving the process of breaking the covalent bonds and the formations thereof, these types of individuals are not of significance. Hence, they require little or no attention, in the context of discovering stereoisomers. In fact, the inevitable evaluations of the individuals of these kinds when using canonical memetic algorithms waste the precious computational time and resources, considering that the calculation of potential energy, force, and frequency of a molecule with a few tens of atoms could be very computationally intensive depending on the fidelity of the molecular model employed.

With all covalent bonds known *a priori*, the connectivity between any pair of atoms in the molecule can be quickly embedded into a general tree structure prior to optimizing the structure of the molecule in the course of searching for its stereoisomers. This one-time embedding then allows all subsequent operators, under the guidance of the constructed tree structure, to produce meaningful – or at least close to meaningful – individuals. Pseudocode in Fig. 2 describes how the tree should be constructed. A reference set, in addition to the tree, is also constructed. This reference set are pointers to the tree nodes which are neither the root nor any leaf node. Crossover and mutation points can then be very quickly selected from the reference set, allowing some tree traversals to be avoided which would otherwise be more computationally demanding. For illustration purposes, Fig. 3 portrays the constructed tree structure and the corresponding reference set of a glutamic acid molecule when atom C3 is selected as a pivot.

4.1.1. Tree-based initialization

The random initialization procedure, as discussed in the previous section, simply randomize each chromosome of every individual in the initial population within the range defined by some lower and upper bounds. In the molecular context,

```

Select a pivot atom and assign it as the root node;
Identify its neighboring atoms and assign them as its children;
En-queue all these children nodes to Q;
Define an empty reference set RS;
While Q is not empty
  De-queue a from Q;
  Identify the neighboring atoms of a and assign them as the children of a unless it is the parent node of a;
  If a has at least one child node
    Put a in RS;
    En-queue all children nodes of a to Q;
  End if
End while

```

Fig. 2. Pseudocode of the construction of a tree representative of a covalently-bonded molecule.

these chromosomes correspond to the atomic coordinates of the molecule of interest. Randomization of the chromosomes, therefore, would easily violate the covalent bonding within the molecule. Two atoms supposedly connected by a covalent bond could end up being far apart from each other as though the covalent bond is broken after a random initialization, and vice versa.

Inspired by the naturally-occurring rotamers, we propose in this subsection a tree-based initialization procedure. Rotamers are basically a subset of stereoisomers in which the isomers can be interconverted through rotations about some covalent bonds. Therefore, the atomic template of the molecule of interest must first be made available. This can easily be downloaded from any publicly accessible databases [43–45] molecular modelling tools [46]. Guided by the tree representative of the molecule and the corresponding reference set thereof, the tree-based initialization procedure then performs a random rotation about the covalent bond between each atom represented as neither the root nor any leaf node and its parent. In other words, a random angle shall technically rotate the subtree of every atom in the reference set about the covalent bond between that particular atom and its parent. This procedure shall be repeated until the initial population is fully populated. In this way, consequently, there is only a tiny probability of producing nonsensical structures of the molecule as two atoms not connected by a covalent bond get too close to each other, causing steric clash. A small percentage of constitutional isomer candidates of the molecule may also be produced as two atoms not connected by a covalent bond come into contact with each other at the optimal distance for a covalent bond to form. Finally, the largest portion of the initial population should intuitively be populated with the stereoisomer candidates of the molecule.

4.1.2. Tree-based crossover

Crossover in the course of searching for stereoisomers using memetic algorithms aims at exchanging the substructures of the molecule of interest from two selected individuals in the population while still maintaining the integrity of the molecule itself. This implies that the resulting offspring must share the same sets of atoms and bonds with their parents. The use of the canonical one-point or two-point crossover operator is often irrelevant since the molecule of interest is generally not a linear sequence of pairs of covalently-bonded atoms. The use of the other conventional crossover operators discussed in the previous section, unfortunately, could also produce rather easily the offspring that resemble none of their parents from the molecular structure perspective. Blending the atomic coordinates of one parent with those of the other parent does not necessarily produce offspring that have the good traits of both parents. For instance, a substructure of the first parent might have already been optimal while the corresponding substructure in the second parent is not yet optimized. Hence, blending the first with the second parent's atomic coordinates of this substructure would result in an unoptimized substructure that intuitively could also violate the covalent bonding within the molecule by producing impossible bond lengths, improbable bond angles, or steric clashes between any two atoms.

In nature, it is known that similar structures of the functional groups, such as the amine and the carboxylic-acid groups of the amino acids, are often observed across many different chemical compounds at their stable conformations. Motivated by this, we propose in this subsection a tree-based crossover operator that exchanges a randomly selected substructure of one parent with the corresponding substructure in the other parent without altering their respective atomic coordinates. This, as a matter of fact, resembles the canonical one-point crossover except for its linear nature; due to which it is possible to have unexchanged substructures within the exchanged substructure as if there were multiple crossover points. The proposed nature-inspired tree-based crossover operator first selects randomly two parent individuals. Following the pseudocode in Fig. 4 to align the parent individuals by calculating the optimal rotation matrix that minimizes their RMSD, as adapted from the Kabsch algorithm [47], the crossover operator then randomly selects one atom from the reference set as the crossover point. All atoms in the subtree where the selected atom is the root node are then exchanged between the parent individuals. An illustrative example with atom C5 selected as the crossover point is portrayed in Fig. 5. Due to the alignment, there exists a minute probability of producing nonsensical structures of the molecule as two atoms not connected by a covalent bond get too close to each other, causing steric clash. A small percentage of constitutional isomer candidates of the molecule may

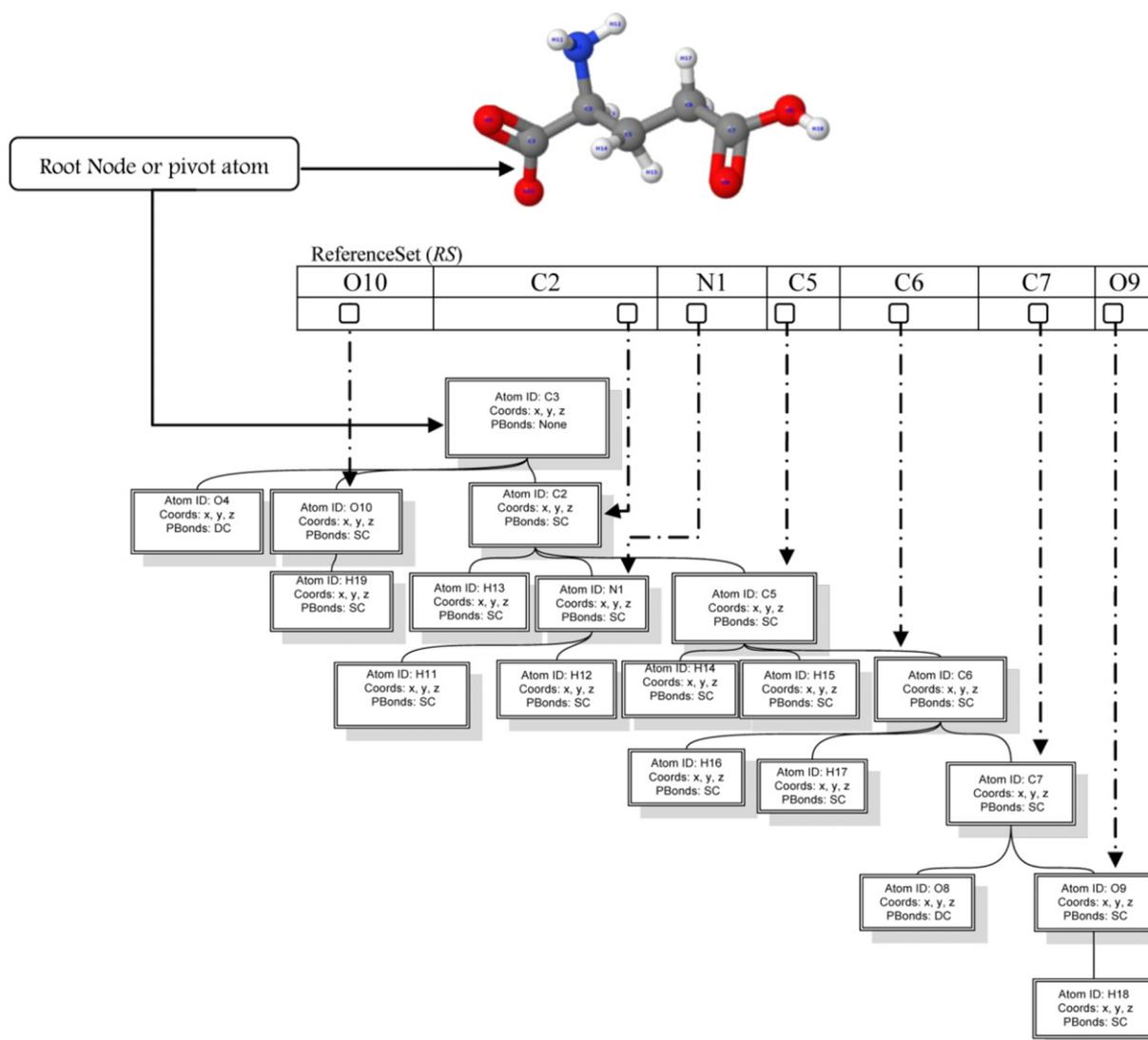


Fig. 3. Example of the tree representative of glutamic acid and the corresponding reference set thereof.

also be produced as two atoms not connected by a covalent bond come into contact with each other at the optimal distance for a covalent bond to form. However, the largest portion of the produced offspring are the stereoisomer candidates of the molecule.

4.1.3. Tree-based mutation

In the course of searching for stereoisomers using memetic algorithms, the use of conventional mutation operators discussed in the previous section can be thought of as a form of local perturbations to individual atoms in some random directions for some random distances. Undoubtedly, this could easily cause impossible bond lengths, improbable bond angles, or the steric clashes between any atom pairs, thereby violating the original covalent bonding of the molecule of interest. Many of the mutated offspring are therefore of negligible importance in the course of searching for stereoisomers, wasting precious computational time and resources performing the CPU-intensive potential energy calculations. A mutation operator that is capable of local perturbations to individual atoms in the feasible directions within reasonable distances is thus of great significance.

Motivated by the bond-stretching mechanism in molecular dynamics simulations due to the atomic-level vibrations in addition to the concept of rotamers introduced earlier, we propose in this subsection a tree-based mutation operator capable of inducing local perturbations to some covalent bonds through rotation and stretching. For rotation, the same technique used in the tree-based initialization also applies here except that only a few – instead of all – atoms in the reference set would be selected to undergo this rotational mutation. For stretching, all atoms but one that corresponds to the root node can be probabilistically selected to undergo mutation. Depending on the type of the selected atoms and their parents, different

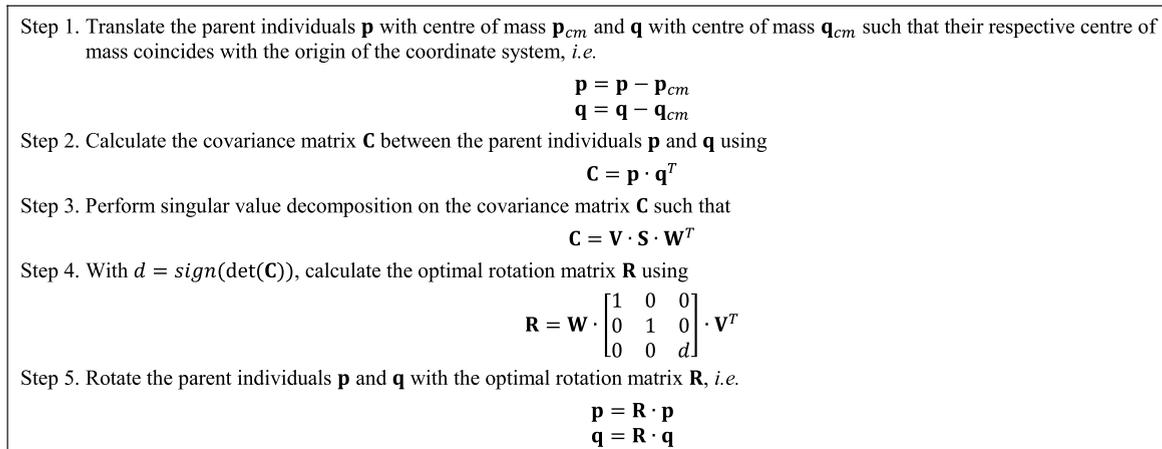


Fig. 4. Pseudocode of the alignment of parent individuals by means of the optimal rotation matrix.

Table 1
Bond length constraints.

Bond type	Minimum bond length (Å)	Maximum bond length (Å)
O–H	0.95	1.15
C–H	0.95	1.15
N–H	1.06	1.12
N–C	1.47	2.10
O–C	1.43	2.15
C–C	1.20	1.54

intervals govern the possible amount of translations in the directions of the covalent bonds between these atoms and the parents. Table 1 tabulates the bond length constraints, as considered by domain experts, of several types of atom pairs that will be considered in this paper, particularly in the empirical study in the next section. For illustration purposes, Fig. 6 portrays both the rotational and translational mutations. In practice, the choice between these two types of mutations is assigned randomly, putting more emphasis on the rotational mutation with a 95%-to-5% probability over the translational mutation.

Due to the local perturbation nature of mutation, there exists an extremely low probability of producing nonsensical structures of the molecule of interest. It is highly unlikely that two atoms not connected by a covalent bond would get too close to each other, thereby causing steric clash, after the mutation. As with the tree-based initialization operator, which performs rotations about covalent bonds, a small percentage of constitutional isomer candidates of the molecule of interest is anticipated. Two atoms not connected by a covalent bond may come into contact with each other at an optimal distance for a covalent bond to form. Nevertheless, the largest portion of the mutated offspring would still be the stereoisomer candidates of the molecule of interest.

4.2. Tree-structured covalent-bond-driven molecular memetic algorithm

With the details of the proposed nature-inspired tree-based evolutionary operators elaborated in the previous subsection, Fig. 7 shall then present the flowchart of the proposed tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA). Embedding domain knowledge prior to commencing the search for stereoisomers in the form of tree-structured connectivity information of the covalent bonding of the molecule of interest, the proposed molecular memetic algorithm is expected to perform more efficiently than its conventional counterparts. Directed by the tree-represented covalent bonds in populating the initial population as well as producing offspring through crossover and mutation, TCM-MA is anticipated to spend most of its effort exploring regions of the search space that contain potential stereoisomer candidates with some reservations for exploring regions of the search space that contain potential constitutional isomer candidates. The TCM-MA might spend little or no effort exploring regions of the search space that contain only nonsensical structures of the molecule of interest.

5. Results and discussions

To assess the efficacy of the proposed algorithm, an empirical study was performed using glutamic acid as the example molecule of interest. The potential energy as well as its gradients and eigenvalues were calculated using an ab-initio approach at the Hartree–Fock level of approximation with STO-3G basis set, which is available in the Gaussian 09 software

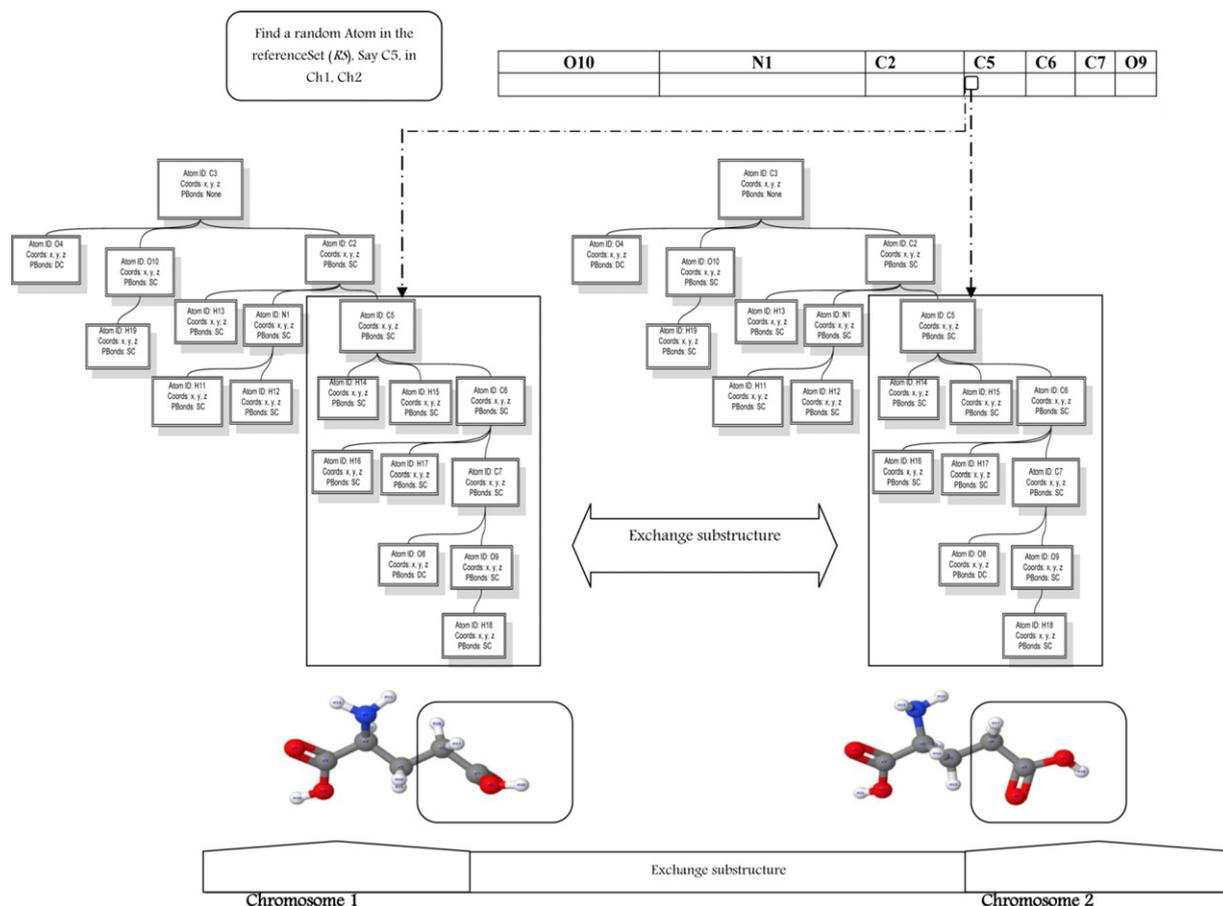


Fig. 5. Illustrative example of tree-based crossover.

Table 2
Parameter settings.

Parameter	Value
Population size	100
Maximum generation	50
Crossover probability	0.5
Mutation probability	0.5
ε	0.0001 Hartree
γ	0.04
λ	0.00035

Table 3
Performance measures.

Performance measure	Description
Success rate	The percentage of algorithm runs that successfully discover at least one stereoisomer
Discovered stereoisomers	The average number of discovered stereoisomers among successful algorithm runs
Gaussian calls	The average number of Gaussian calls for energy-related evaluations among successful algorithm runs

package. The algorithmic parameter settings used are as tabulated in Table 2. The comprehensive set of criteria presented in Table 3 then serves as the various performance measures used to validate the algorithms considered.

In this section, relative performance of the proposed nature-inspired tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA) as compared to the state-of-the-art Sequential Niching Memetic Algorithm (SNMA) [48] and the conventional Stochastic Multi-start Local Search (SMLS) [49] is presented. Table 4 quantifies the average performance over ten independent runs in terms of success rate, number of discovered stereoisomers, and number of Gaussian calls during the course of finding the stereoisomers. With a 100% success rate, the TCM-MA is shown to outperform the other

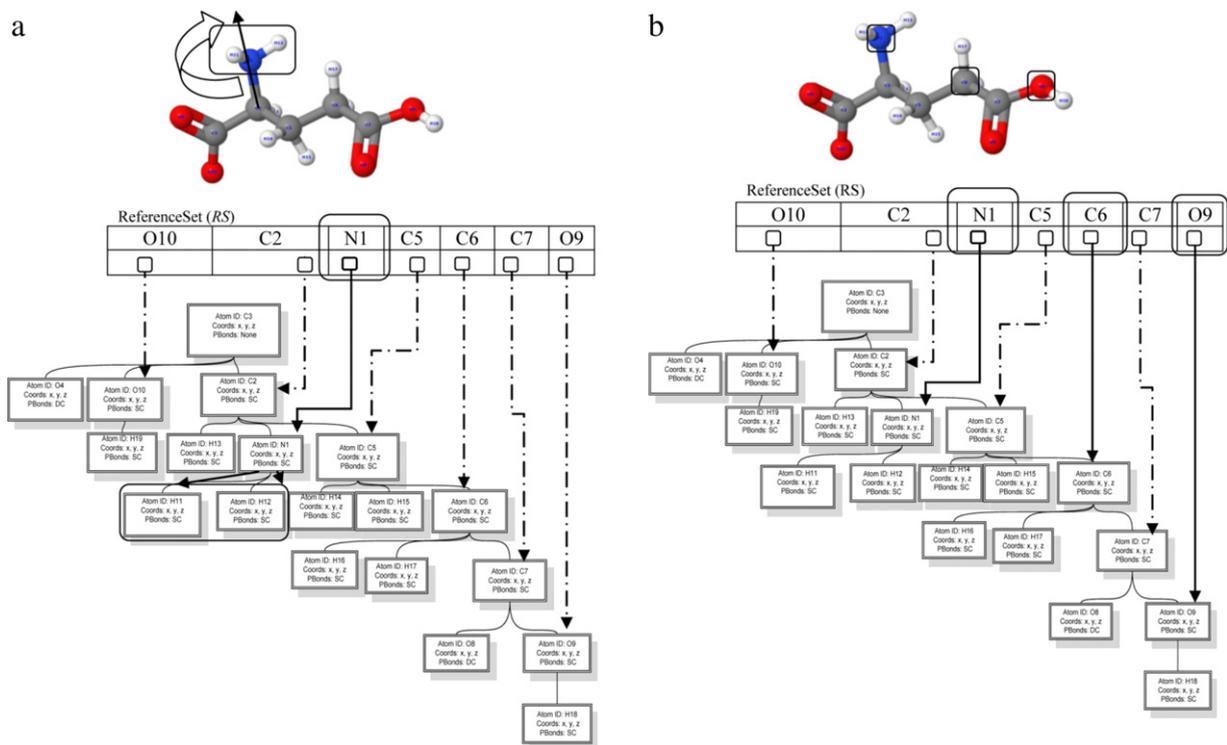


Fig. 6. Illustrative example of tree-based mutation: (a) rotational; (b) translational.

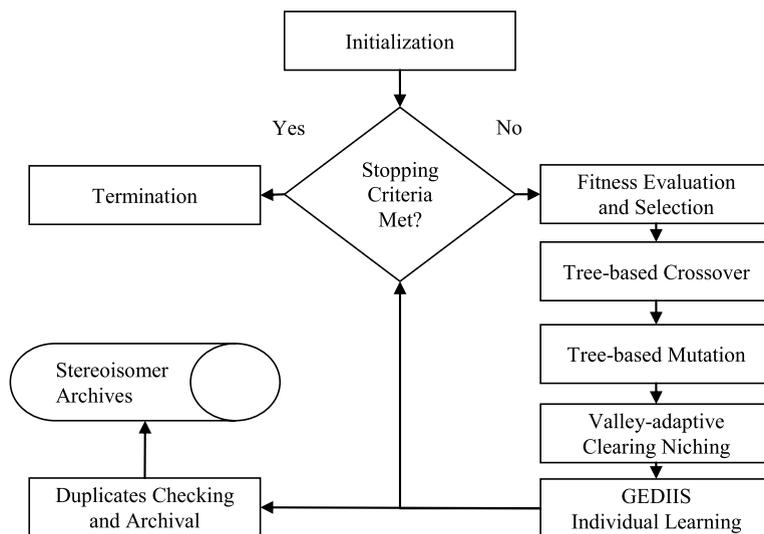


Fig. 7. Flowchart of the proposed tree-structured covalent-bond-driven molecular memetic algorithm.

Table 4
Performance of different algorithms over 10 independent runs.

	TCM-MA	SNMA	SMLS
Success rate	100%	60%	40%
Discovered stereoisomers	47 ± 7	5 ± 2	3 ± 1
Gaussian calls	6,771 ± 574	30,185 ± 1,119	40,000 ± 2789

two algorithms significantly, discovering as many as up to 16 times more stereoisomers while requiring only as little as up to one-sixth of the total computational budget incurred by the other two competing algorithms.

Table 5
Performance of different initialization operators over 10 independent memetic algorithm runs.

	Tree-based initialization	Random initialization
Success rate	100%	2%
Discovered stereoisomers	46 ± 2	3 ± 0
Gaussian calls	12,639 ± 272	13,295 ± 423

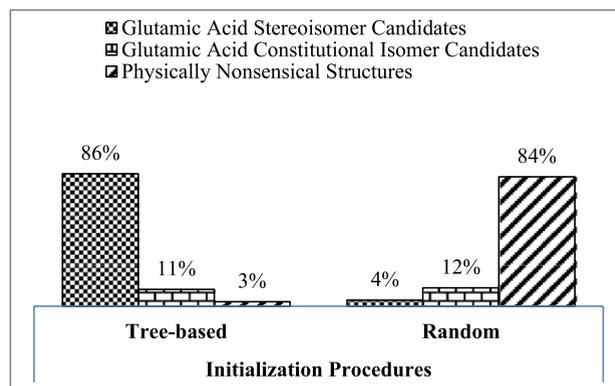


Fig. 8. Relative percentage of glutamic acid stereoisomer candidates, constitutional isomer candidates, and physically nonsensical structures generated by different initialization operators.

To provide detailed insights into the high efficacy of the proposed algorithm, the following three subsections shall provide further empirical analyses of the individual contribution of each nature-inspired tree-based operator of the TCM-MA. In each subsection, the proposed operator shall be compared with other commonly-used operators. Comparisons of algorithmic performances under the use of different operators will first be presented and discussed. The number of potential candidates for stereoisomers as well as constitutional isomers and the number of meaningless or impossible isomers produced by the various operators will also be reported to complete the discussions.

5.1. Empirical study on initialization operators

In this section, the PBX crossover and the Makinen–Periaux–Toivanen (M–P–T) mutation operators are employed and coupled with the valley-adaptive clearing method as the niching and the GEDIIS method as the individual-learning operator. Performance of the molecular memetic algorithm when initialized using the proposed nature-inspired tree-based initialization as opposed to when initialized randomly is then measured over ten independent runs. Note that atomic coordinates are randomly sampled from the rectangular search space in the range of -5 and 5 Å when initialized randomly. The success rate as well as the average number of stereoisomers discovered and Gaussian calls over the successful memetic algorithm runs as tabulated in Table 5 clearly indicates that the tree-based initialization procedure outperforms the random initialization procedure in all aspects. The percentage of potential isomers generated during the initialization as shown in Fig. 8 intuitively explains the extremely low success rate and number of stereoisomers discovered when the random initialization procedure was in use. Thus, it can be concluded that the proposed nature-inspired tree-based initialization operator contributes significantly to the success of the molecular memetic algorithm in finding more stereoisomers within a tighter computational budget.

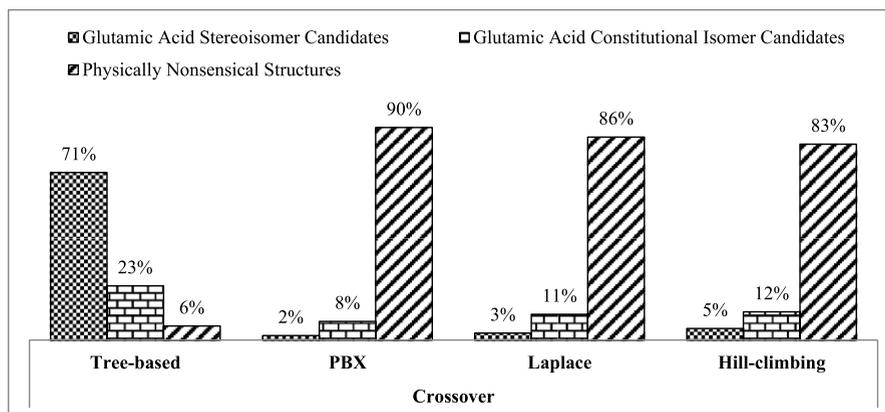
5.2. Empirical study on crossover operators

To study the individual contribution of the proposed nature-inspired tree-based crossover operator to the efficacy of the proposed covalent-bond-driven molecular memetic algorithm, ten independent runs of the memetic algorithm were performed. This is done by employing the proposed tree-based initialization operator, the valley-adaptive clearing method as niching operator, the GEDIIS method as individual-learning operator, and four different crossover operators while setting the mutation probability to zero so as to isolate the reproductive contribution that is solely from the crossover operator alone. It should be noted that tree-based crossover would require the initial population to be generated by means of tree-based initialization. In Table 6, the summary of the average performance of the ten independent memetic algorithm runs under different crossover operators is presented, from which it is observable that the key to 100% success rate is actually the tree-based initialization coupled with the tree-based crossover operator. Without a tree-based crossover, the result in Table 5 implies that the mutation probability cannot be zero in order to achieve a 100% success rate. In terms of number of stereoisomers discovered, there is a slight decrease compared to the figure in Table 5 despite the use of the tree-based initialization, probably due to the lack of mutation such that the memetic algorithm would fail to locate a small number of

Table 6

Performance of different crossover operators over 10 independent memetic algorithm runs.

	Tree-based crossover	PBX crossover	Laplace crossover	Hill-climbing crossover
Success rate	100%	30%	30%	40%
Discovered stereoisomers	38 ± 4	5 ± 4	7 ± 3	9 ± 6
Gaussian calls	9,129 ± 3,046	12,355 ± 612	12,355 ± 612	17,693 ± 852

**Fig. 9.** Relative percentage of glutamic acid stereoisomer candidates, constitutional isomer candidates, and physically nonsensical structures generated by different crossover operators.

promising neighboring valleys. However, an interesting observation is seen in terms of number of Gaussian calls. The use of tree-based crossover seems to have complemented the strength of the tree-based initialization, reducing the number of Gaussian calls as compared to the use of other crossover operators.

To explain this phenomenon, the number of potential stereoisomers as well as constitutional isomers and the number of meaningless or impossible isomers produced by the four crossover operators during the course of the ten memetic algorithm runs were computed. As summarized by Fig. 9, the tree-based crossover produces the largest number of potential stereoisomers as well as constitutional isomers, and thus, is expected to accelerate the convergence of the individual-learning operator. It is also observed that contrary to the other three crossover operators, the tree-based crossover produces more potential candidates for stereoisomers than constitutional isomers. Upon producing meaningless structures or the potential candidates for constitutional isomers, additional Gaussian calls would generally be expected to rectify these structures before they can be refined so as to converge to some precise stereoisomers. A smaller number of Gaussian calls when employing the tree-based crossover as opposed to when using the other three crossover operators, therefore, is expected as the result.

5.3. Empirical study on mutation operators

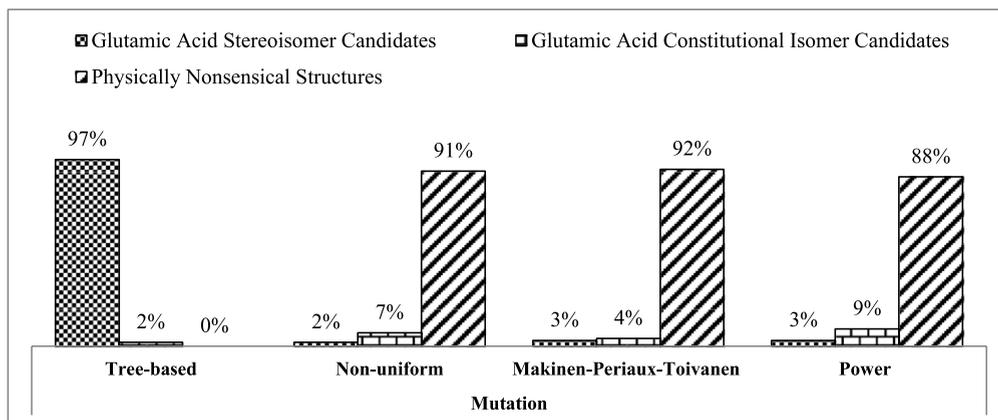
Similar to the previous subsection, ten independent runs of the covalent-bond-driven molecular memetic algorithm – employing the tree-based initialization operator, the valley-adaptive clearing method as niching operator, and the GEDIIS method as individual-learning operator – were performed to study the individual contribution of the proposed nature-inspired tree-based mutation operator to the efficacy of the proposed memetic algorithm. Likewise, it should also be noted that tree-based mutation would require the initial population to be generated by means of tree-based initialization. In contrast to the previous subsection, however, the crossover probability is the one that was set to zero while experimenting with the four different mutation operators so as to isolate the reproductive contribution solely from the mutation operator alone. In Table 7, a summary of the average performance of the ten independent memetic algorithm runs under different mutation operators is presented, from which it is observable that the use of tree-based initialization is indeed evident to 100% success rate. Unlike in the previous subsection where a 100% success rate were not achieved whenever the tree-based crossover was not in use, the less destructive perturbation effect of mutation as compared to that of the crossover must have participated in the 100% success rate achieved in this section regardless of the mutation operator in use.

In terms of number of stereoisomers discovered, it can be deduced from Tables 5–7 that setting the crossover probability to zero deteriorates the performance significantly. This is understandable as with mutation alone, the memetic algorithm would only be able to locate a small number of neighboring valleys, therefore having only a slim probability of finding new stereoisomers. Like in the previous subsection, however, an interesting observation is seen in terms of number of Gaussian calls. Nonetheless, only when the tree-based mutation operator was used will the number of Gaussian calls be reduced. To explain this phenomenon, the number of potential stereoisomers as well as constitutional isomers and the number of meaningless or impossible isomers produced by the four mutation operators in the course of the ten memetic algorithm

Table 7

Performance of different mutation operators over 10 independent memetic algorithm runs.

	Tree-based mutation	Non-uniform mutation	M–P–T mutation	Power mutation
Success rate	100%	100%	100%	100%
Discovered stereoisomers	17 ± 1	2 ± 1	5 ± 0	4 ± 1
Gaussian calls	7,056 ± 565	37,862 ± 532	39,825 ± 185	41,894 ± 3,307

**Fig. 10.** Relative percentage of glutamic acid stereoisomer candidates, constitutional isomer candidates, and physically nonsensical structures generated by different mutation operators.

runs were computed. As summarized by Fig. 10, the tree-based mutation operator produces the largest number of potential stereoisomers, and thus, is expected to greatly accelerate the convergence of the individual-learning operator. Furthermore, the nearly 100% potential stereoisomers generated by using the tree-based mutation benefits the memetic algorithm in that less effort would be spent on rectifying the potential constitutional isomers or the meaningless structures, thus requiring a smaller number of Gaussian calls.

6. Conclusion and future works

An efficient tree-structured covalent-bond-driven molecular memetic algorithm (TCM-MA) is proposed in this paper, employing the novel nature-inspired tree-based evolutionary operators. The proposed algorithm demonstrated a high efficacy when an empirical study with glutamic acid as the example molecule of interest was conducted. The algorithm successfully discovered more stereoisomers of the molecule within a tighter computational budget as compared to the state-of-the-art Sequential Niching Memetic Algorithm (SNMA) [48] and the conventional Stochastic Multi-start Local Search (SMLS) [49]. An in-depth experimentation with each of the three nature-inspired tree-based evolutionary operators shows nonsensical structures of the glutamic acid molecule occupy only a tiny fraction of the produced offspring individuals. As intended, the glutamic acid stereoisomer candidates become the dominant offspring individuals with a small percentage of the constitutional isomer candidates of the molecule of interest. The use of these evolutionary operators in the TCM-MA – preceded with construction of the tree representative of the connectivity information about covalent bonding in the molecule of interest so as to restrain exploration of the search space to its most promising regions – then allows the proposed molecular memetic algorithm to eliminate unnecessary efforts of searching in the infeasible solution space. Coupling the proposed algorithm with the Feasibility Structure Modeling paradigm [22,50,51] in the near future, it is envisaged that improved efficiency of the molecular memetic algorithm shall be attained, hence liberating much of the precious computational time and resources for some other purposes including the search for the stereoisomers of other molecules.

Glutamic acid, a major neurotransmitter in the central nervous system that plays a key role in brain functions and neurological disorders [8], is one of many biomolecules essential for life. Recent studies show that different functions may be assumed as these biomolecules adopt various low-energy stable conformations, allowing selective or preferential interactions with different systems [4–7]. Henceforth, the identification of these conformations is of significant importance when it comes to revealing the structure–function relationship of the biomolecules. Challenged by the difficulty in using experimental methods for stereoisomer identification, the computational approach provides a more practical alternative. Generalization of the proposed algorithm in the near future to cyclic ring-inclusive molecules with the automatic selection of the pivot atom would thus allow stereoisomer identification of molecules like beta amyloid, which is a biomolecule of great interest in the research of Alzheimer's disease. This shows that the proposed algorithm with nature-inspired tree-based evolutionary operators indeed paves the way for computational stereoisomer identification of important biomolecules.

References

- [1] D.S. Goodsell, *The Machinery of Life*, Springer, 1997.
- [2] J. Chen, S.J. Swamidass, Y. Dou, J. Bruand, P. Baldi, ChEMDB: a public database of small molecules and related cheminformatics resources, *Bioinformatics* 21 (2005) 4133–4139.
- [3] S.L. Schreiber, Organic chemistry: molecular diversity by design, *Nature* 457 (2009) 153–154.
- [4] N. Nassar, J. Cancelas, J. Zheng, D.A. Williams, Y. Zheng, Structure-function based design of small molecule inhibitors targeting rho family gtpases, *Current Topics in Medicinal Chemistry* 6 (2006) 1109–1116.
- [5] J. Lu, Z. Ma, J.C. Hsieh, C.W. Fan, B. Chen, J.C. Longgood, N.S. Williams, J.F. Amatruda, L. Lum, C. Chen, Structure-activity relationship studies of small-molecule inhibitors of wnt response, *Bioorganic & Medicinal Chemistry Letters* 19 (2009) 3825–3827.
- [6] Ceroni, F. Costa, P. Frasconi, Classification of small molecules by two- and three-dimensional decomposition kernels, *Bioinformatics* 23 (2007) 2038–2045.
- [7] S.J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, P. Baldi, Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity, *Bioinformatics* 21 (Suppl. 1) (2005) i359–i368.
- [8] R.M. Sapolsky, *Biology and Human Behavior: The Neurological Origins of Individuality*, The Teaching Company, 1996.
- [9] F. Tellier, F. Acher, I. Brabet, J.P. Pin, J. Bockaert, R. Azerad, Synthesis of conformationally-constrained stereospecific analogs of glutamic acid as antagonists of metabotropic receptors, *Bioorganic & Medicinal Chemistry Letters* 5 (1995) 2627–2632.
- [10] K. Shimamoto, M. Ishida, H. Shinozaki, Y. Ohfune, Synthesis of four diastereomeric L-2-(carboxycyclopropyl)glycines. Conformationally constrained L-glutamate analogs, *The Journal of Organic Chemistry* 56 (1991) 4167–4176.
- [11] R.D. Paz, S. Tardito, M. Atzori, K.Y. Tseng, Glutamatergic dysfunction in schizophrenia: from basic neuroscience to clinical psychopharmacology, *European Neuropsychopharmacology* 18 (2008) 773–786.
- [12] D'Orlando, B.T. Fellay, B. Schwaller, V. Salicio, A. Bloc, V. Gotzos, M.R. Celio, Calretinin and calbindin D-28k delay the onset of cell death after excitotoxic stimulation in transfected P19 cells, *Brain Research* 909 (2001) 145–158.
- [13] ChEBI, Glutamic acid. Available: <https://www.ebi.ac.uk/chebi/searchId.do?chebiId=18237>, 2011.
- [14] Brooks 3rd, J. Onuchic, D. Wales, Statistical thermodynamics. Taking a walk on a landscape, *Science* 293 (2001) 612–613.
- [15] Wales, *Energy Landscapes: [with Applications to Clusters, Biomolecules and Glasses]*, Cambridge University Press, 2003.
- [16] H. Soh, Y.S. Ong, Q.C. Nguyen, Q.H. Nguyen, M.S. Habibullah, T. Hung, J.-L. Kuo, Discovering unique, low-energy pure water isomers: memetic exploration, optimization and landscape analysis, *IEEE Transactions on Evolutionary Computation* 14 (2010) 419–437.
- [17] P.J. Ballester, Ultrafast shape recognition: method and applications, *Future Medicinal Chemistry* 3 (2011) 65–78.
- [18] P.J. Ballester, P.W. Finn, W.G. Richards, Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology, *Journal of Molecular Graphics & Modelling* 27 (2009) 836–845.
- [19] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes, *Journal of Computational Chemistry* 28 (2007) 1711–1723.
- [20] Y.S. Ong, M.H. Lim, X.S. Chen, Research frontier: memetic computation – past, present & future, *IEEE Computational Intelligence Magazine* 5 (2010) 24–36.
- [21] X.S. Chen, Y.S. Ong, M.H. Lim, K.C. Tan, A multi-facet survey on memetic computation, *IEEE Transactions on Evolutionary Computation* 15 (2011) 591–607.
- [22] S.D. Handoko, C.K. Kwok, Y.S. Ong, Feasibility structure modeling: an effective chaperon for constrained memetic algorithms, *IEEE Transactions on Evolutionary Computation* 14 (2010) 740–758.
- [23] M.N. Le, Y.S. Ong, Y. Jin, B. Sendhoff, Lamarckian memetic algorithms: local optimum and connectivity structure analysis, *Memetic Computing Journal* 1 (2009) 175–190.
- [24] J. Chia, C. Goh, K. Tan, V. Shim, Memetic informed evolutionary optimization via data mining, *Memetic Computing* 3 (2011) 73–87.
- [25] W. Jakob, A general cost-benefit-based adaptation framework for multimeme algorithms, *Memetic Computing* 2 (2010) 201–218.
- [26] Z. Zexuan, J. Sen, J. Zhen, Towards a memetic feature selection paradigm [application notes], *Computational Intelligence Magazine, IEEE* 5 (2010) 41–53.
- [27] P. Moscato, *On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms*, California Institute of Technology, 1989.
- [28] Q.C. Nguyen, Y.S. Ong, M.H. Lim, A probabilistic memetic framework, *IEEE Transactions on Evolutionary Computation* 13 (2009) 604–623.
- [29] R.E. Bellman, *Dynamic Programming*, Courier Dover Publications, 2003.
- [30] L.J. Eshelman, J.D. Schaffer, Real-coded genetic algorithms and interval-schemata, *Foundation of Genetic Algorithms* 2 (1993) 187–202.
- [31] K. Deep, M. Thakur, A new crossover operator for real coded genetic algorithms, *Applied Mathematics and Computation* 188 (2007) 895–911.
- [32] M. Lozano, F. Herrera, N. Krasnogor, D. Molina, Real-coded memetic algorithms with crossover hill-climbing, *Evolutionary Computation* 12 (2004) 273–302.
- [33] R.A.E. Mäkinen, J. Periaux, J. Toivanen, Multidisciplinary shape optimization in aerodynamics and electromagnetics using genetic algorithms, *International Journal for Numerical Methods in Fluids* 30 (1998) 149–159.
- [34] Z. Michalewicz, T. Logan, S. Swaminathan, Evolutionary operators for continuous convex parameter space, in: Presented at the Third Annual Conference on Evolutionary Programming, River Edge, New Jersey, USA, 1994.
- [35] Petrowski, A clearing procedure as a niching method for genetic algorithms, in: Presented at the IEEE International Conference on Evolutionary Computation, Nayoya University, Japan, 1996.
- [36] Singh, K. Deb, Comparison of multi-modal optimization algorithms based on evolutionary algorithms, in: Presented at the Eighth Annual Conference on Genetic and Evolutionary Computation, Seattle, Washington, USA, 2006.
- [37] M.M.H. Ellabaan, Y.S. Ong, Valley-adaptive clearing scheme for multimodal optimization evolutionary search, in: Presented at the Ninth International Conference on Intelligent Systems Design and Applications, Pisa, Italy, 2009.
- [38] L. Steinmetz, Using the method of steepest descent, *Industrial & Engineering Chemistry* 58 (1966) 33–39.
- [39] O. Peitgen, *Newton's Method and Dynamical Systems*, Springer, 1989.
- [40] O. Farkas, H.B. Schlegel, Geometry optimization methods for modeling large molecules, *Journal of Molecular Structure-Theochem* 666 (2003) 31–39.
- [41] C. Mauro, R.J. Loucks, J. Balakrishnan, A simplified eigenvector-following technique for locating transition points in an energy landscape, *The Journal of Physical Chemistry A* 109 (2005) 9578–9583.
- [42] X. Li, M.J. Frisch, Energy-represented direct inversion in the iterative subspace within a hybrid geometry optimization method, *Journal of Chemical Theory and Computation* 2 (2006) 835–839.
- [43] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic Acids Research* 36 (2008) D344–D350.
- [44] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Research* 39 (2011) 1–8.
- [45] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Poon, K. Banco, C. Mak, V. Neveu, Y. Djombou, R. Eisner, A.C. Guo, D.S. Wishart, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, *Nucleic Acids Research* 39 (2011) D1035–D1041.
- [46] Avogadro: an open-source molecular builder and visualization tool. Available: <http://avogadro.openmolecules.net/>.
- [47] W. Kabsch, A solution of the best rotation to relate two sets of vectors, *Acta Crystallographica* 32 (1976) 922–923.
- [48] E. Vitela, O. Castanos, A real-coded niching memetic algorithm for continuous multimodal function optimization, in: Presented at the IEEE World Congress on Computational Intelligence, Hong Kong, 2008.
- [49] R. Martí, Multi-start methods, *Handbook of Metaheuristics* (2003) 355–368.
- [50] S.D. Handoko, C.K. Kwok, Y.S. Ong, Classification-assisted memetic algorithms for equality-constrained optimization problems, *Lecture Notes in Computer Science* 5866 (2009) 391–400.
- [51] S.D. Handoko, C.K. Kwok, Y.S. Ong, J. Chan, Classification-assisted memetic algorithms for solving optimization problems with restricted equality constraint function mapping, in: Presented at the IEEE Congress on Evolutionary Computation, New Orleans, Louisiana, USA, 2011.