Determining the Conformational Flexibility of Disaccharides with an Adaptive Differential Evolution Approach

Alfeu Uzai Tavares
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
alfeu.uzai@inf.ufrgs.br

Márcio Dorn
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
mdorn@inf.ufrgs.br

***Abstract*—Determining the three-dimensional structure of a biomolecule is of great importance for understanding its biological functions. Carbohydrates are an essential and diverse family of biomolecules that exerts many functions. The structure of carbohydrates, especially those of smaller sizes such as the disaccharides, are difficult to be studied by experimental methods. One alternative is the use of computational methods such as Molecular Dynamics to infer the conformation of disaccharides. However, Molecular Dynamics simulations are computationally expensive due to the many force field evaluations and atom motions required to simulate the studied molecular system. In this work, we represent the disaccharide structure prediction problem as an optimization problem. We propose a metaheuristic based on the Success-History Adaptive Differential Evolution algorithm to find the best conformation, using the GROMOS force field energy function with 53A6$_{\text{GLYC}}$ parameters as the objective function. We tested the proposed method against different disaccharide structures. A comparative analysis between the results achieved by our method and those obtained by Molecular Dynamics simulations were performed. The results show that the proposed method achieved satisfactory results in terms of accuracy and with a reduced computational time when compared to the usual approach of Molecular Dynamics.**

***Index Terms*—structural biology, disaccharides, optimization, metaheuristic, differential evolution**

I. Introduction

The determination and study of the three-dimensional structures of biomolecules are of great importance for the understanding of the biological functions they perform [1], [2]. Structural properties are fundamental for several areas of study, whether for the understanding of the functioning of cellular mechanisms, treatment of diseases, and the development of new drugs [3]. The available techniques to study molecular structures are divided into two main classes: experimental and computational. Within the computational methods, Molecular Dynamics (MD) is compelling regarding some aspects, such as its lower financial cost, greater accessibility, and the ability to deal with systems that can be challenging to study using experimental techniques. On the other hand, one of its negative aspects is related to the high computational cost due to numerous evaluations of its energy function and movement of the system atoms. A MD simulation must be done for a time interval long enough in order to allow a proper study of the system [4].

Carbohydrates are biomolecules composed mostly of carbon atoms, oxygen, and hydrogen. They are essential molecules that supply non-photosynthetic cells with energy [5]. Besides, they also have structural functions and act as an energetic reserve. There is a great variety of forms where structures can appear in the form of simple monomers up to large polymers with hundreds of monosaccharide units [6]. Several works seek to study their functions and determine their structures. Many of them use computational techniques, especially MD, due to the inherent difficulty of studying small saccharides through experimental techniques [7], [8]. In this paper, we explore the development and application of a Differential Evolution (DE) [9] metaheuristic as an alternative method to Molecular Dynamics. We seek to obtain results similar to those obtained by a MD simulation, with reduced running time.

II. Preliminaries

A. Carbohydrates

Carbohydrates (or saccharides) are the most abundant and diverse biomolecules in nature. They perform numerous biological functions and are formed by atoms of carbon, oxygen, and hydrogen, mostly following the formula $\mathrm{C}_m(\mathrm{H}_2\mathrm{O})_n$. They can be classified regarding the degree of polymerization of the molecule [6]: (i) *monosaccharides*: basic unit that makes up the other carbohydrates; (ii) *disaccharides*: formed by the bonding of two monosaccharides; (iii) *oligosaccharides*: formed by three up to twenty monosaccharides; (iv) *polysaccharides*: formed by more than twenty monosaccharides.

There is a vast diversity of carbohydrates due to a large number of distinct monosaccharide units and the many possibilities of bonding between them. Such diversity leads carbohydrates to play numerous roles in organisms, such as energy storage, structural support of cells (cellulose and chitin), support of the RNA chain (ribose), coenzymes components, and interaction with proteins (glycoconjugates) [10].

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

Monosaccharides: Are the simplest carbohydrates, serving as the basic units that make up the other saccharides. They follow the formula $(CH_2O)_x$, and can be classified according to the number of carbon atoms such as trioses ($x = 3$), tetroses ($x = 4$), pentoses ($x = 5$), hexoses ($x = 6$) and heptoses ($x = 7$). Pentoses and hexoses are also called furanose and pyranose, respectively [10]. Monosaccharide carbon atoms are numbered incrementally, starting from the carbon belonging to the carbonyl group (C=O) (Figure 1). The rotational configuration of the most distant carbonyl determines two isomers named D and L [11]. In hexoses, the D isomers are observed in nature at a higher frequency. Monosaccharides are found in acyclic or cyclic form, the latter being the most common for pentoses and hexoses. In the cyclic configuration, the position of the hydroxyl (OH) from the carbon $C1$ defines two anomers, α, and β. The anomer is defined depending on the relative position between the hydroxyl from the $C1$ atom and the CH_2OH group attached to $C5$. If both groups are on the same side of the plane defined by the ring, the anomer β is configured, otherwise α if they are on opposite sides (Table I). In this work, the studied monosaccharides belong to the D-aldohexose group, with cyclic chain, in its two anomeric forms α and β, totalizing 16 units of monosaccharides (Table I).

TABLE I: D-aldohexose monosaccharides studied in this work, α (top half) and β anomers (bottom half).

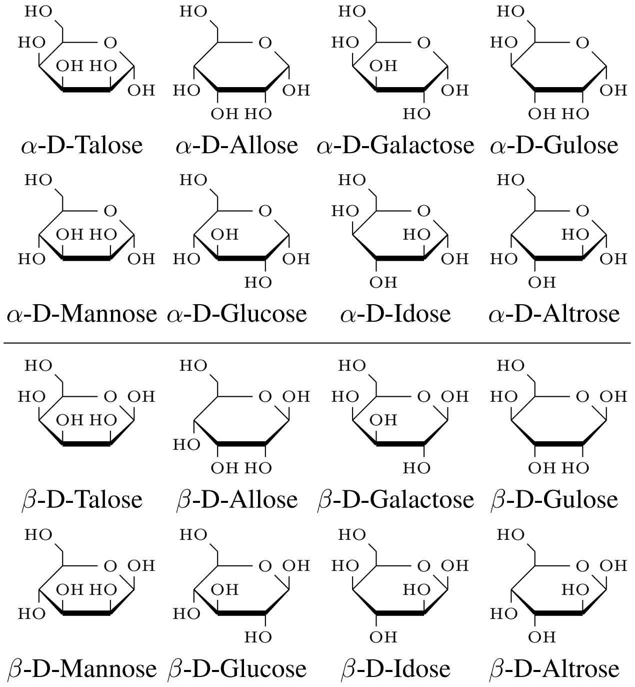
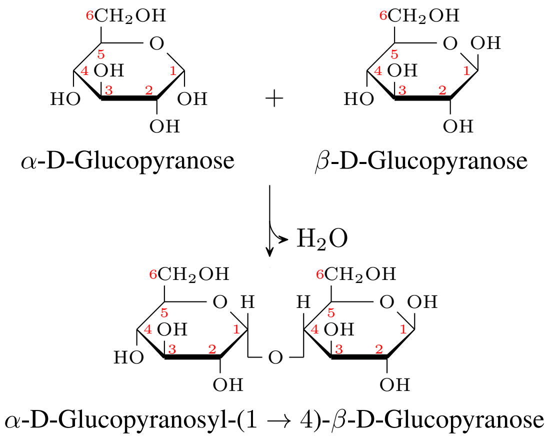


Fig. 1: Schematic representation of the formation of a disaccharide from the glycosidic bond between two monosaccharide units. Carbon numbers shown in red.



Disaccharides: Are composed by two monosaccharides units joined by a glycosidic bond [10] [6]. The bond occurs between two hydroxyl groups, one from each monosaccharide, releasing an water molecule in the process (Figure 1). In this work, the studied disaccharides bonds are established between the hydroxyl group of the carbon $C1$ from the first monosaccharide unit and the hydroxyl group attached to the CX carbon, with $X \in \{2, 3, 4, 6\}$, from the second monosaccharide, in a total of 1024 distinct disaccharides. The established bond gives rise to two or three new dihedral angles of great importance for the structural description of the disaccharide [6]. In $1 \rightarrow X$ bonds, with $X \in \{2, 3, 4, 6\}$, the two new dihedral angles are ϕ and ψ. The $1 \rightarrow 6$ bonds also have a third new dihedral ω. A dihedral angle can be defined by four consecutively bonded atoms. The angles ϕ, ψ and ω in the disaccharides uses the following convention to choose the atoms i, j, k and l (Figure 2c):

$$\phi : O5 - C1 - OX - CX$$
$$\psi : C1 - OX - CX - C(X-1)$$
$$\omega : O6 - C6 - C5 - C4$$

B. *Differential Evolution*

Differential Evolution (DE) is a population-based metaheuristic for numerical optimization, designed to operate on real-valued functions ($f : \mathbb{R}^D \rightarrow \mathbb{R}$) [9]. The DE fits into a sub-class of metaheuristics named evolutionary algorithms. Such algorithms were conceived inspired by biological processes, especially in the theory of evolution. They usually undergo the phases of initialization, crossover, mutation, and selection. The DE algorithm operates over a population of real-valued vectors $P = \{x_i \in \mathbb{R}^D \mid i \in [1, NP]\}$, representing the candidate solutions for the objective function f. After the initialization phase (e.g., randomly sampling the vectors), the algorithm executes the mutation, crossover and selection phases iteratively, until reaching some stop condition.

At each generation G, a mutant vector $v_{i,G}$ is generated for each vector $x_{i,G} \in P$. Several mutation strategies are available, and the following equations exemplifies those most commonly found in the literature:

- **rand/1**:

$$v_{i,G} = x_{r_1,G} + F(x_{r_2,G} - x_{r_3,G}) \tag{1}$$

- **best/1**:

$$v_{i,G} = x_{best,G} + F(x_{r_1,G} - x_{r_2,G}) \tag{2}$$

- **current-to-best/1**:

$$\begin{aligned} v_{i,G} = x_{i,G} &+ F_1(x_{best,G} - x_{i,G}) \\ &+ F_2(x_{r_1,G} - x_{r_2,G}) \end{aligned} \tag{3}$$

The indexes r_1, r_2 and r_3 are randomly sampled from $\{1, ..., NP\} \setminus \{i\}$ and differ from each other (for each vector). The index x_{best} refers to the current vector in the population with the best objective value (e.g., the lowest value, assuming minimization of f). The mutation parameter $F \in [0.0, 1.0]$ controls the mutation scale. The crossover operation between the respective pair of current and mutant vectors results in the trial vector $u_{i,G}$. The most usual crossover operator used in DE is the binomial (or uniform) crossover, where:

$$u_{i,G,j} = \begin{cases} v_{i,G,j} & \text{, if rand(0, 1)} \leqslant CR \text{ or } j = j_{rand}, \\ x_{i,G,j} & \text{, otherwise.} \end{cases} \tag{4}$$

For each dimension $j \in [1, D]$ a value is inherited from the mutant vector with CR probability. A randomly selected dimension (j_{rand}) ensures that always at least one value is inherited from the mutant vector. The trial vectors are then evaluated by the objective function, and replaces their respective current vectors $x_{i,G+1}$ in the next population if $f(u_{i,G}) < f(x_{i,G})$ (assuming minimization).

III. Proposed Method

A. Disaccharides Representation

A vector of real values must represent the three-dimensional structure of the disaccharides in order to be suited for the DE algorithm. The representation also must be able to describe the various conformational configurations of the disaccharides. For such, the intramolecular parameters (geometrical properties) of the molecule are used [4]. They provide a straightforward representation of the disaccharide structure and can also be easily restricted to valid values with real intervals, unlike the explicit use of atomic coordinates.

Five types of intramolecular parameters are used: bond lengths, bond angles, dihedral angles, improper dihedrals, and ring puckering coordinates (Fig. 2, Fig. 3). The first four parameters are also used in the energy function, detailed in the next subsection. The ring puckering coordinates are based on the work by Cremer and Pople [12].

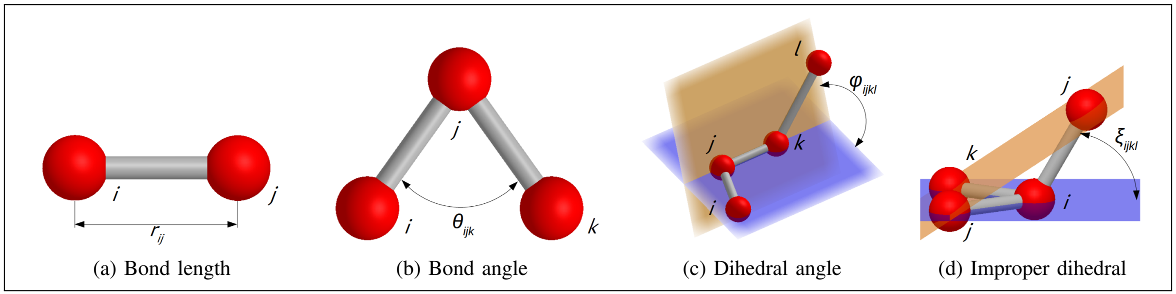
In the following descriptions of each parameter, atoms are identified by the letters i, j, k and l, whereas $i - j$ denotes that i and j are bonded atoms. The notation $\vec{\mathbf{r}}_{\mathbf{ij}}$ represents the vector between the atoms i and j, and $\hat{\mathbf{r}}$ is an unit vector.

Bond lengths (Fig. 3-a): one parameter for each $i - j$ bond is used, except for those belonging to the rings. When setting a bond parameter only the length of the $\vec{\mathbf{r}}_{\mathbf{ij}}$ vector is changed, i.e., the other atoms connected to j are also moved, keeping the other distances and angles of the molecule. The length values are restricted to the interval $[|\vec{\mathbf{r}}_{\mathbf{ij_{initial}}}| - 0.02, |\vec{\mathbf{r}}_{\mathbf{ij_{initial}}}| + 0.02]$ (values in nm).

Bond angles (Fig. 3-b): it is used one parameter for each angle defined by the three atoms $i - j - k$, except for those belonging to the rings. Setting an angle value only changes the angle between the vectors $\vec{\mathbf{r}}_{\mathbf{ji}}$ and $\vec{\mathbf{r}}_{\mathbf{jk}}$. For this purpose, the coordinates of the k atom and the other atoms connected to it are rotated along the axis defined by $\hat{\mathbf{r}}_{\mathbf{ji}} \times \hat{\mathbf{r}}_{\mathbf{jk}}$, with j as the origin. Bond angles values are restricted to the interval $[0.0, \pi]$.

Dihedral angles (Fig. 3-c): a proper dihedral angle is defined by the four atoms $i - j - k - l$, and consists in the angle between the two planes that contains the atoms $i - j - k$ and $j - k - l$. To set the dihedral angle the atom l and other atoms connected to it are rotated along the axis $\hat{\mathbf{r}}_{\mathbf{jk}}$, with j as the origin. A parameter is used for every dihedral angle that doesn't changes the ring, and the values are restricted to the interval $[-\pi, \pi]$.

Improper dihedrals (Fig. 3-d): in the improper dihedrals the four atoms are chosen in a specific order to represent the chirality of a tetrahedral configuration, where the atoms j, k and l are bonded to the central atom i. A parameter is used for each carbon atom in the rings (i atom of the improper), with values restricted to the interval $[-\frac{2\pi}{9}, \frac{2\pi}{9}]$ (same as $[-40°, 40°]$ in degrees).



(a) Bond length (b) Bond angle (c) Dihedral angle (d) Improper dihedral

Fig. 2: Intramolecular parameters used in the disaccharides representation.

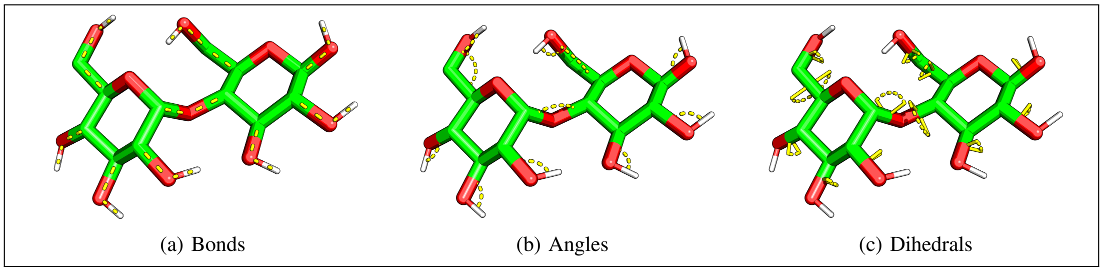


Fig. 3: Example with bond lengths (a), bond angles (b) and dihedral angles (c) used to represent the disaccharides conformations.

Ring puckering: the ring puckering coordinates of Cremer and Pople [12] allows the description of N-atoms rings conformations with a reduced number of parameters. For six atom rings, as in the studied disaccharides, three coordinates define the ring puckering: the puckering amplitude Q and the torsion angles θ and ϕ. With the purpose to define the disaccharides rings conformation given these coordinates, the equations and procedures originally proposed in [12] for measurements can be used as follows: *(1)* the geometrical mean plane of the ring is computed; *(2)* the current displacements z_i of each ring atom to this plane is then computed; *(3)* using the expression for z_i in function of Q, θ and ϕ the correct displacements z_i^{new} are calculated in function of the puckering coordinates (Equation 5); *(4)* finally, the ring atoms are displaced by $z_i^{new} - z_i$ along the mean plane's normal, establishing the ring puckering defined by the given coordinates. The atoms connected to the ring atom are also displaced to maintain the other geometrical properties of the molecule. Puckering coordinates are restricted to $Q \in [0.0, 0.08]$ (nm), $\theta \in [0.0, \pi]$ and $\phi \in [-\pi, \pi]$.

$$\begin{gathered} z_i = \frac{1}{\sqrt{3}} q_2 \cos(\phi + 2\pi(i-1)/3) + \frac{1}{\sqrt{6}} q_3 (-1)^{i-1} \\ q_2 = Qsin(\theta) \qquad q_3 = Qcos(\theta) \qquad i \in [1, 6] \end{gathered} \tag{5}$$

It should be noted that the configuration of some geometrical parameters may still affect others as a side effect. Using the following order when setting the disaccharide structure based on the representation vector values, those unwanted side effects are removed: first, the rings puckering is configured, followed by the bond lengths, bond angles, improper dihedrals, and finally, the dihedral angles.

B. Objective Function

The structural stability of a molecule is related to lower values of its free energy [13]. Taking this into account, we use as the objective function to model the optimization problem the energy function from the force field GROMOS [14], with the parameterization 53A6$_{\mathrm{GLYC}}$ [15]. This parameter set was specially developed to better represent the monosaccharide units considered in this work.

The general form of the GROMOS energy function is:

$$V(r; s) = V^{phys}(r; s) + V^{special}(r; s) \tag{6}$$

$$V^{phys}(r; s) = V^{bon}(r; s) + V^{nbon}(r; s) \tag{7}$$

Where the argument r is the atom coordinates of the system, and s is the associated parameterization. The term $V^{special}$ is used to restrain specific properties of the molecules (such as bond lengths), and will not be used since the restrictions (real intervals) violations of the representation vector will be handled by the metaheuristic.

The term V^{phys} is subdivided into V^{bon}, related to the intramolecular potentials, and V^{nbon}, which represents the electrostatic potentials between non-bonded atoms. The V^{bon} terms calculates the intramolecular potentials of the bonds lengths, bonds angles, dihedral angles and improper dihedrals, as show bellow in equations 8, respectively.

$$\begin{gathered} V^{bond}(r; K_b, b_0) = \sum_{n=1}^{N_b} \frac{1}{4} K_{b_n} [{b_n}^2 - {b_{0_n}}^2]^2 \\ V^{angle}(r; K_\theta, \theta_0) = \sum_{n=1}^{N_\theta} \frac{1}{2} K_{\theta_n} [\cos\theta_n - \cos\theta_{0_n}]^2 \\ V^{trig}(r; K_\varphi, \delta, m) = \sum_{n=1}^{N_\varphi} K_{\varphi_n} [1 + \cos(\delta_n)\cos(m_n \varphi_n)] \\ V^{har}(r; K_\xi, \xi_0) = \sum_{n=1}^{N_\xi} \frac{1}{2} K_{\xi_n} [\xi_n - \xi_{0_n}]^2 \end{gathered} \tag{8}$$

Three terms composes the electrostatic potential V^{nbon} between pairs of non-bonded atoms: V^{LJ} (Lennard-Jones potential), V^C (Coulombic interactions) and V^{RF} (reaction-field), shown below in the equations 9.

$$\begin{gathered} V^{LJ}(r; C12, C6) = \sum_{pairs\ i,j} \left(\frac{C12_{ij}}{r_{ij}^{12}} - \frac{C6_{ij}}{r_{ij}^{6}} \right) \\ V^C(r; q) = \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1} \frac{1}{r_{ij}} \\ V^{RF}(r; q) = \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1} \frac{-\frac{1}{2}C_{rf} r_{ij}^2}{R_{rf}^3} \end{gathered} \tag{9}$$

The term V^{RF} is not used, since no cutoff radius (R_{rf}) will be used. Then, the final used energy function terms are:

$$F(r; s) = V^{bon}(r; s) + V^{LJ}(r; s) + V^{C}(r; s) \tag{10}$$

The objective function $f_s(x)$ is first initialized with the parameters s for the specific disaccharide being evaluated. To evaluate a representation vector x, the coordinates of the disaccharide are first set accordingly to the intramolecular parameters. The objective value of f_s is the total energy computed with Equation 10.

C. Proposed Search Strategy

Determining the conformation of a disaccharide can be seen as an optimization problem of the type $f : \mathbb{R}^D \to \mathbb{R}$, where intervals constrain the values of $x \in \mathbb{R}^D$. Many metaheuristics have been developed to address this class of problems, with the DE algorithm being one of them that has been showing an excellent performance. Furthermore, several works were developed seeking to improve the canonical DE, especially by the use of adaptation mechanisms for the parameters F and CR. Among these works, the Success-History based Adaptive Differential Evolution (SHADE) [16] algorithm and its variants (such as the L-SHADE algorithm [17]) have obtained excellent results in several CEC competitions [18]. Using an adaptive parameter mechanism will not only provide better and more robust performance but will also eliminate the need to manually tune the parameters on a case-by-case basis (for a total of 1024 different disaccharides).

The DE algorithm used in this paper is based mostly on the SHADE algorithm, and also uses the population size reduction from the L-SHADE. In the SHADE algorithm, there are three main modifications:

1) **External archive:** an aditional population of vectors A is used to maintain the population diversity. Vectors $x_{i,G}$ that are replaced by a better trial vector $u_{i,G}$ in the selection phase are assigned to this population. The size of A is kept equal to NP, and randomly selected vectors are removed when $|A| > NP$.

2) **current-to-pbest/1 mutation strategy:** this mutation strategy is similar to the *current-to-best* mutation (Equation 3), with the x_{best} vector actually being selected ramdomly from the p% best vectors of the current population (for each mutant), and the x_{r2} vectors selected from $P \cup A$. These modifications aims to control the greedines of the mutation, achieving a better balance between exploration and exploitation.

3) **Parameters adaptation:** the parameters F_i and CR_i are individually generated for each $x_{i,G}$ in every generation. F_i values are sampled from a Cauchy distribution (Equation 12), being resampled when lesser than 0.0 and truncated to 1.0 if greater than 1.0. CR_i values are drawn from a normal distribution (Equation 13) and truncated to the [0.0, 1.0] interval. The mean values of each distribution are randomly selected from the historical memories M_F and M_{CR} for each vector (Equation 11).

$$r = rand(1, H) \tag{11}$$

$$F_i = cauchy(M_{F,r}, 0.1) \tag{12}$$

$$CR_i = normal(M_{CR,r}, 0.1) \tag{13}$$

Each memory consists of H entries with the mean values of the F_i and CR_i parameters used to generate the successfull $u_{i,G}$ of a generation. During the selection phase, when a trial vector is selected to replace its parent $x_{i,G}$, the associated F_i and CR_i values are stored in the vectors S_F and S_{CR}. The improvement rate $\Delta f_i = |f(u_{i,G}) - f(x_{i,G})|$ is also recorded. Then, at the end of a generation, the current memory entry $k \in [1, H]$ is updated. The new F mean is calculated with a weighted Lehmer mean (Equation 15), and the new CR value by a weighted arithmetic mean (Equation 16). Both equations uses the improvement rates as the weighting factors, emphasizing parameters associated with a greater improvement (Equation 14).

$$w_j = \frac{\Delta f_j}{\sum_{j=1}^{|\Delta f|} \Delta f_j} \tag{14}$$

$$M_{F,k} = \frac{\sum_{j=1}^{|S_F|} w_j S_{F,j}^2}{\sum_{j=1}^{|S_F|} w_j S_{F,j}} \tag{15}$$

$$M_{CR,k} = \sum_{j=1}^{|S_{CR}|} w_j S_{CR,j} \tag{16}$$

In the L-SHADE algorithm, the most significant modification is the population size reduction. Starting with NP^{init} vectors, at the end of each generation a new NP value is computed (Equation 17), and the worst solutions from the population are removed until $|P| = NP$. The population size decreases to NP^{min} after MAX_NFE function evaluations, also used as the stop condition.

$$NP_{g+1}(NFE) = round\left[\left(\frac{NP^{min} - NP^{init}}{MAX\_NFE}\right) NFE + NP^{init}\right] \tag{17}$$

In the proposed method, the weights F_1 and F_2 (Equation 3) are adapted independently of each other, whereas in the original SHADE algorithm the same value is used to both mutation parameters.

To repair any eventual boundary violations that can happen when computing the mutant vectors, the repetition of the mutation equation with re-selected x_{r_1} and x_{r_2} vectors is performed until a feasible vector is generated [19]. A pseudo-code of the algorithm is shown next, highlighting the main modifications from the canonical DE.

Algorithm 1: SHADE + POPULATION SIZE REDUCTION

```
Input: $f : \mathbb{R}^D \to \mathbb{R}$, $x^{low}$, $x^{high}$, $NP^{init}$, $NP^{min}$,
       MAX_NFE, H
begin
    $P = \{\mathbf{x}_i | i = 1, ..., NP^{init}\}$, $x_{i,j} \sim U(x_j^{low}, x_j^{high})$
    NP = $NP^{init}$
    $A = \{\}$
    Initialize memories $M_{F_1}$, $M_{F_2}$ and $M_{CR}$, $k = 1$
    NFE = $NP^{init}$
    while NFE + NP $\leqslant$ MAX_NFE do
        for $x_{i,G} \in P$ do
            $F_{1i}, F_{2i}, CR_i = generate(M_{F_1}, M_{F_2}, M_{CR})$
            $v_{i,G}$ = curr-to-pbest$(P, A, x_{i,G}, F_1i, F_2i)$
            Repair boundary violations of $v_{i,G}$
            $u_{i,G} = crossover(x_{i,G}, v_{i,G}, CR_i)$
        end
        $S_{F_1} = \{\}, S_{F_2} = \{\}, S_{CR} = \{\}, \Delta f = \{\}$
        for $x_{i,G} \in P$ do
            if $f(u_{i,G}) < f(x_{i,G})$ then
                Store $F_{1i}$, $F_{2i}$, $CR_i$ and $\Delta f_i$
                Add $x_{i,G}$ to $A$
                $x_{i,g+1} = u_{i,G}$
                NFE++
            end
        end
        $update(k, S_{F_1}, S_{F_2}, S_{CR}, \Delta f)$
        $NP = NP_{g+1}(NFE)$
        Resize $P$ if $|P| > NP$, resize $A$ if $|A| > NP$
    end
end
```

IV. Experiments and Results

A. Previous Work

In our experiments we studied monosaccharides belonging to the D-aldohexose group, with cyclic chain, in its two anomeric forms α and β, totalizing 16 units of monosaccharides (Table I). A previous work studied the same group of disaccharides using Molecular Dynamics [20]. We use this study as a reference to analyze the results obtained by the metaheuristic. It also used the same force field (GROMOS 53A6$_{\text{GLYC}}$), and two main steps where performed: (i) *Metadynamics:* in this step were obtained structures of global and local minimal free energy for each disaccharide, along with a contour map of the mean free energy; (ii) *Molecular Dynamics:* the structures of minimum energy (local and global) obtained in the previous step undergoes a MD simulation to study the stability of the structures and the interconversions between the local and global minima. Only 544 from the total of 1024 disaccharides were studied with Metadynamics, and from those 478 also went to the Molecular Dynamics step. One important difference to mention is the use of explicit water molecules in the Metadynamics/MD, which is not being considered in the proposed method.

B. Experimental Setup

Both the DE and MD were executed in the same machine: *IBM X3650 M5 Server; Intel Xeon E5-2650V4 30 MB, 2 CPUs, 2.2Ghz, 48 cores/threads; 64 GB; Titan X Pascal, 3584 CUDA core, 12 GB GDDR5X*. Since the DE is a stochastic algorithm, it was executed 31 times for each disaccharide. The following parameters were used by the metaheuristic:

- $NP^{init} = 200$
- $NP^{min} = 3$
- $MAX\_NFE = 200000$
- $H = 100$
- ***p*** value used in the *curr-to-pbest* mutation: 0.2
- Initial memories values: $M_{F_1} = 0.01$, $M_{F_2} = 0.9$ and $M_{CR} = 0.01$

NP^{init}, MAX_NFE and the initial values for M_{F_1}, M_{F_2} and M_{CR} where chosen empirically by running the method for a small set of disaccharides, being selected those that provided the lowest best objective value means and standard deviation in 31 runs. NP^{min} is the smallest value possible for the *curr-to-pbest* mutation, and H and p values are the same as reported by the original papers.

C. Results

Regarding the general (ϕ, ψ) distribution, the proposed method was able to achieve results consistent with those obtained by MD simulations. The main common aspects observed by both methods are: *(i)*: concentration of (ϕ, ψ) values in the quadrants centers; *(*ii): influence of the first monosaccharide anomerism, with α anomers occurring mostly in quadrants I and IV (Figure 4) and β anomers in quadrants II and III (Figure 5); *(iii)*: higher occurrence of ψ values near $-\pi$ and π for 1→6 bonds.

To perform a case-by-case comparison, two sets of data are available from the previous work: the (ϕ, ψ) values of minimal free energy (local and global) from the Metadynamics step and the most frequent values after the Molecular Dynamics refinement. To compare these results with those of metaheuristics, the Chebyshev distance between the (ϕ, ψ) values from the DE and the Metadynamics/MD is used (Equation 19) with a small modification to take into account the toroidal nature of the space (Equation 18). DE results with a distance of up to 45°(average radius of the most frequent regions in the MD simulations) will be considered similar to those of the Metadynamics/MD. When more than one pair of angles are available from the Metadynamics/MD results, those that present the smallest distance are used.

$$d(\theta_a, \theta_b) = \begin{cases} |\theta_a - \theta_b|, & \text{if } |\theta_a - \theta_b| \leqslant \pi \\ 2\pi - |\theta_a - \theta_b|, & \text{otherwise} \end{cases} \quad (18)$$

$$D((\phi_a, \psi_a), (\phi_b, \psi_b)) = max(d(\phi_a, \phi_b), d(\psi_a, \psi_b)) \quad (19)$$

A first comparison was done using the best results (lowest energy) from the 31 runs of the metaheuristic againts both the Metadynamics and Molecular Dynamics results (tables II and III). In a second comparison the results from all runs for each disaccharide are considered, being used the pair of angles that presented the shortest distance (tables IV and V). The first column of the tables idintifies the disaccharyde group by bond number and anomerism of the first monosaccharide, followed

Fig. 4: Best (ϕ, ψ) of each disaccharide, α anomers.

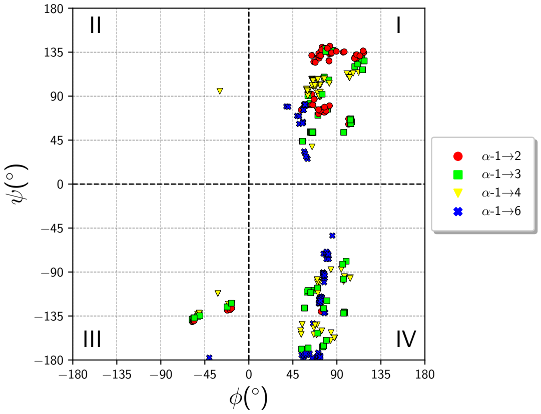


Fig. 5: Best (ϕ, ψ) of each disaccharide, β anomers.

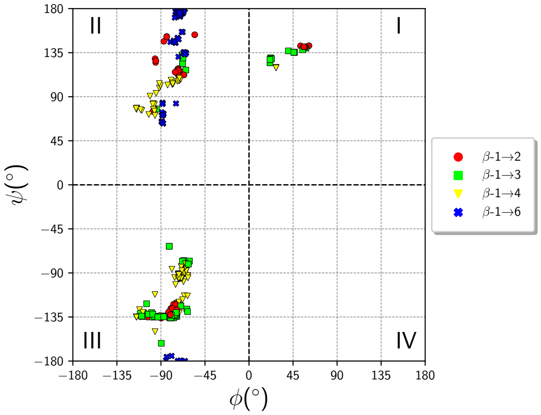


TABLE II: DE best runs × Metadynamics

Group	#Total	$\#\phi_{dist} \leqslant 45°$	$\#\psi_{dist} \leqslant 45°$	$\#(\phi,\psi)_{dist} \leqslant 45°$
α-1→2	67	63 (94.03%)	52 (77.61%)	52 (77.61%)
α-1→3	67	56 (83.58%)	38 (56.72%)	32 (47.76%)
α-1→4	67	60 (89.55%)	35 (52.24%)	33 (49.25%)
α-1→6	76	72 (94.74%)	52 (68.42%)	48 (63.16%)
All α	277	251 (90.61%)	177 (63.90%)	165 (59.57%)
β-1→2	71	63 (88.73%)	25 (35.21%)	17 (23.94%)
β-1→3	77	63 (81.82%)	29 (37.66%)	23 (29.87%)
β-1→4	67	64 (95.52%)	45 (67.16%)	45 (67.16%)
β-1→6	62	50 (80.65%)	38 (61.29%)	29 (46.77%)
All β	277	240 (86.64%)	137 (49.46%)	114 (41.16%)
All	554	491 (88.63%)	314 (56.68%)	279 (50.36%)

TABLE III: DE best runs × Molecular Dynamics

Group	#Total	$\#\phi_{dist} \leqslant 45°$	$\#\psi_{dist} \leqslant 45°$	$\#(\phi,\psi)_{dist} \leqslant 45°$
α-1→2	67	57 (85.07%)	57 (85.07%)	56 (83.58%)
α-1→3	64	53 (82.81%)	26 (40.62%)	25 (39.06%)
α-1→4	62	55 (88.71%)	14 (22.58%)	8 (12.90%)
α-1→6	67	61 (91.04%)	24 (35.82%)	24 (35.82%)
All α	260	226 (86.92%)	121 (46.54%)	113 (43.46%)
β-1→2	70	54 (77.14%)	26 (37.14%)	15 (21.43%)
β-1→3	53	41 (77.36%)	10 (18.87%)	1 (1.89%)
β-1→4	46	42 (91.30%)	31 (67.39%)	30 (65.22%)
β-1→6	49	32 (65.31%)	43 (87.76%)	32 (65.31%)
All β	218	169 (77.52%)	110 (50.46%)	78 (35.78%)
All	478	395 (82.64%)	231 (48.33%)	191 (39.96%)

by the total number of disaccharides in the second column. In the following columns is shown the number of cases that the proposed method achieved a distance up to 45°, fist comparing ϕ and ψ independently (Equation 18) and finally the total distance (Equation 19).

In both comparisons, the DE achieved a better correspondence with the Metadynamics results rather than those of the Molecular Dynamics. Also, when considering the results from all runs rather than only the best run, a higher similarity is observed between the techniques. When comparing the ϕ and ψ angles independently, ϕ values obtained a better correspondence than ψ values. Such observations provide an indication that in cases where the proposed method achieved an unsatisfactory result are likely due to the convergence to a local minima. As reported by the previous work, the multiple global and local minima of a disaccharide are usually scattered along oposite sides in the ψ axis. As an example, is shown by Figure 6 regions of lowest energy explored during all 31 runs of a case that the DE result did not achieve the expected outcome (with $\phi = -67.15°$and $\psi = 110.14°$) in quadrant II, despite having intensively explored such region.

Concerning the running time the proposed method greatly outperformed the Metadynamics/MD, taking around 7s per run (3min 37s for all 31 runs of a single disaccharide; single core) compared to 3h (CPU + GPU with GROMACS [21]). This gain is mostly due to the reduced number of function evaluations, of $200,000$ in the metaheuristic whereas for the Metadynimics/MD with a total simulation time of $110ns$ and $dt = 0.001ps$, $110,000,000$ evaluations are performed. Also, the presence of explicity water molecules in the MD makes its function evaluations more costly. In addition, the metaheuristic doesn't need to compute the forces in each atom and their motions, since the structures are changed by the search method.

Fig. 6: (ϕ, ψ) regions of lowest energy (at most 5% higher than the best value) during the 31 runs of the metaheuristic for the disaccharide β-D-Allose-(1→3)-β-D-Idose. Local minima explored in quadrant II, but runs converged in quadrant III.

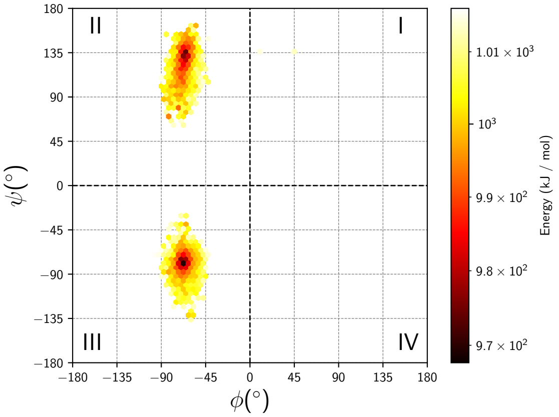


TABLE IV: DE all runs × Metadynamics

Group	#Total	#$\phi_{dist} \leqslant 45^\circ$	#$\psi_{dist} \leqslant 45^\circ$	#$(\phi,\psi)_{dist} \leqslant 45^\circ$
α-1→2	67	67 (100.00%)	63 (94.03%)	63 (94.03%)
α-1→3	67	65 (97.01%)	55 (82.09%)	54 (80.60%)
α-1→4	67	62 (92.54%)	52 (77.61%)	49 (73.13%)
α-1→6	76	75 (98.68%)	64 (84.21%)	63 (82.89%)
All α	277	269 (97.11%)	234 (84.48%)	229 (82.67%)
β-1→2	71	71 (100.00%)	54 (76.06%)	54 (76.06%)
β-1→3	77	69 (89.61%)	47 (61.04%)	40 (51.95%)
β-1→4	67	64 (95.52%)	57 (85.07%)	56 (83.58%)
β-1→6	62	50 (80.65%)	42 (67.74%)	33 (53.23%)
All β	277	254 (91.70%)	200 (72.20%)	183 (66.06%)
All	554	523 (94.40%)	434 (78.34%)	412 (74.37%)

TABLE V: DE all runs × Molecular Dynamics

Group	#Total	#$\phi_{dist} \leqslant 45^\circ$	#$\psi_{dist} \leqslant 45^\circ$	#$(\phi,\psi)_{dist} \leqslant 45^\circ$
α-1→2	67	67 (100.00%)	66 (98.51%)	66 (98.51%)
α-1→3	64	63 (98.44%)	54 (84.38%)	53 (82.81%)
α-1→4	62	55 (88.71%)	28 (45.16%)	21 (33.87%)
α-1→6	67	65 (97.01%)	57 (85.07%)	55 (82.09%)
All α	260	250 (96.15%)	205 (78.85%)	195 (75.00%)
β-1→2	70	64 (91.43%)	49 (70.00%)	48 (68.57%)
β-1→3	53	45 (84.91%)	16 (30.19%)	11 (20.75%)
β-1→4	46	45 (97.83%)	44 (95.65%)	44 (95.65%)
β-1→6	49	35 (71.43%)	46 (93.88%)	35 (71.43%)
All β	218	189 (86.70%)	155 (71.10%)	138 (63.30%)
All	478	439 (91.84%)	360 (75.31%)	333 (69.67%)

V. Conclusions

In the presented work, a disaccharides representation sufficiently descriptive of their structural conformations and also suitable to be used with a DE algorithm was successfully defined and implemented. The used objective function reasonably models the problem, but it seems necessary the addition of some solvation term (implicit or explicit) to achieve results closer to those of the Metadynamics and MD. When comparing the results obtained by the proposed method, the general conformational preference of disaccharides is consistent with those reported by the previous work. However, in a case-by-case comparison, there are some differences, mostly due to the multimodal and dynamic nature of the problem.

Despite the metaheuristics results not being a direct match to those obtained by the MD they are of a reasonable quality and obtained quickly. A possible application of the technique would be the initialization of saccharide structures that are part of a more extensive molecular system in a MD simulation, adding a negligible overhead to the total run time. Besides the possible convergence to local minima, the main drawbacks of the proposed method are the lack of explicit water molecules and the loss of physical meaning from the algorithm execution, alongside with any sort of trajectory analysis that could be done in a MD simulation.

In future works, the addition of a solvation term to the energy function and the implementation of mechanisms to allow the metaheuristics to better handle the multimodal nature of the problem are likely to improve the proposed method results. Another possible extension is the addition of more monosaccharide units and the capability of running the proposed technique to optimize more complex saccharides (polysaccharides) instead of only disaccharides.

Acknowledgment

This work was supported by grants from MCT/CNPq [*311611/2018-4*], FAPERGS [*19/25510001906-8*], Alexander von Humboldt-Stiftung (AvH) [*BRA 1190826 HFST* CAPES-P] - Germany, and was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. We gratefully acknowledge the support of the NVIDIA Corporation.

References

[1] N. R. Council, *Opportunities in Biology*. Washington, DC: The National Academies Press, 1989. [Online]. Available: https://www.nap.edu/catalog/742/opportunities-in-biology

[2] A. Shurki and A. Warshel, "Structure/function correlations of proteins using mm, qm/mm, and related approaches: Methods, concepts, pitfalls, and current progress," *Advances in protein chemistry*, vol. 66, pp. 249–313, 02 2003.

[3] C. Wong and A. McCammon, "Protein simulation and drug design," *Advances in protein chemistry*, vol. 66, pp. 87–121, 02 2003.

[4] A. R. Leach, "Empirical force field models: Molecular mechanics," p. 165, 2008.

[5] N. Sharon, "Complex carbohydrates - their chemistry, biosynthesis and functions," *Carbohydrate-Peptide linkages*, pp. 65–83, 01 1975.

[6] V. Rao, *Conformation of Carbohydrates*, 08 2019.

[7] F. V. Toukach and V. P. Ananikov, "Recent advances in computational predictions of nmr parameters for the structure elucidation of carbohydrates: methods and limitations," *Chem. Soc. Rev.*, vol. 42, pp. 8376–8415, 2013.

[8] A. Imberty and S. Perez, "Structure, conformation, and dynamics of bioactive oligosaccharides: Theoretical approaches and experimental validations," *Chemical reviews*, vol. 100, pp. 4567–4588, 2001.

[9] R. Storn and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341–359, 01 1997.

[10] D. Nelson and M. Cox, "Carbohydrates and glycobiology," *Lehninger Principles of Biochemistry*, pp. 311–318, 01 2004.

[11] D. Voet and J. G. Voet, *Biochemistry*. USA: John Wiley &#38; Sons, Inc., 2006.

[12] D. Cremer and J. Pople, "General definition of ring puckering coordinates," *Journal of The American Chemical Society - J AM CHEM SOC*, vol. 97, 03 1975.

[13] W. Gunsteren, "Biomolecular modeling: Goals, problems, perspectives," *ChemInform*, vol. 37, 10 2006.

[14] C. Oostenbrink, A. Villa, A. Mark, and W. van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6," *Journal of computational chemistry*, vol. 25, pp. 1656–76, 10 2004.

[15] L. Pol-fachin *et al.*, "Gromos 53a6 glyc, an improved gromos force field for hexopyranose-based carbohydrates," *J. Chem. Theory Comput.*, vol. 8, p. 4681–4690, 2012.

[16] R. Tanabe and A. Fukunaga, "Success-history based parameter adaptation for differential evolution," 06 2013, pp. 71–78.

[17] R. Tanabe and A. S. Fukunaga, "Improving the search performance of shade using linear population size reduction," in *2014 IEEE Congress on Evolutionary Computation (CEC)*, July 2014, pp. 1658–1665.

[18] D. Molina, F. Moreno-Garcia, and F. Herrera, "Analysis among winners of different ieee cec competitions on real-parameters optimization: Is there always improvement?" 06 2017, pp. 805–812.

[19] V. Kreischer, T. Tavares Magalhães, H. Barbosa, and E. Krempser, "Evaluation of bound constraints handling methods in differential evolution using the cec2017 benchmark," 10 2017.

[20] R. B. Toscan, "Desenvolvimento de abordagem automatizada para construção de banco de dados de preferências conformacionais de carboidratos," Bachelor's Thesis, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 2016, [Online; accessed 29-January-2020]. [Online]. Available: https://lume.ufrgs.br/handle/10183/170127

[21] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. Mark, and H. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, pp. 1701–18, 12 2005.