Computer-Aided Design and Applications

http://www.cadanda.com

eBGF: An Enhanced Geometric Hierarchical Representation for Protein Modeling and Rapid Self-Collision Detection

Athina N. Brintaki[1] and Susana K. Lai-Yuen[2]

[1]University of South Florida, abrintak@mail.usf.edu
[2]University of South Florida, laiyuen@eng.usf.edu

ABSTRACT

The identification of biological molecules is vital for the design and manufacturing of pharmaceutical drugs, nanomaterials and nanoscale products. Understanding molecules' behavior and motion is of fundamental importance for exploring potential nanoscale and drug designs prior to actual fabrication. A major challenge in modeling flexible molecules is the exponential explosion in computational complexity due to the large number of degrees of freedom within a molecular structure. In this paper, a new improved biologically-inspired geometric methodology called enhanced BioGeoFilter (eBGF) is proposed to simplify the representation of macromolecules such as proteins for studying their behavior and motion. The eBGF algorithm can be used as an effective identification tool for molecular feasibility that minimizes molecular conformational search time and speeds collision detection for bio-nano computer-aided design applications.

Keywords: nanoscale design, bio-CAD, protein modeling, self-collision detection.
DOI: 10.3722/cadaps.2009.625-638

1. INTRODUCTION

Bionanotechnology is a key emerging scientific and technological area of nanotechnology that aims to identify and assemble biological molecules to design bionanoscale products with unimaginable applications in every aspect of life. The ability to control and manipulate biological molecules can lead to the development of new pharmaceutical drugs, atomically structured materials, and precise nanoscale devices with new capabilities for diagnosis and treatment of diseases. However, to achieve bionanotechnology, it is essential for researchers to be able to visualize the interactions between the molecular components in real-time during the design stage so that fully functional bionanoscale products can be designed and evaluated prior to actual fabrication.

A main key for enabling the visualization of the molecular components is the understanding and modeling of biological molecules and their interactions in real-time. In recent years, new approaches have been investigated to facilitate nanoscale design by providing real-time force feedback using haptic devices [3], [8], [10]-[13], [17], [18], [20]. Haptic devices are electromechanical devices that exert forces on users giving them the illusion of touching something in the virtual world. These devices have been used to manipulate virtual molecules and to feel the forces as the molecules interact with each other providing an essential design and visualization tool as shown in Fig. 1.

However, current methods using haptics either model molecules as rigid bodies or are limited to local molecular motions and short periods of simulation time. Modeling molecules as rigid bodies can simplify the calculation of forces but does not represent the molecular interactions realistically. The

main difficulty of modeling flexible molecules (or molecular conformations) lies on the exponential explosion in computational complexity as the size of the molecule increases and a large number of degrees of freedom are considered to represent the molecule's flexibility.

Fig. 1: Haptic graphical user interface.

Modeling the behavior of flexible molecules consists of searching for all possible molecular conformations to identify a stable molecular state, which corresponds to a feasible molecular conformation with low internal energy value. Geometrically, a molecular conformation can be considered feasible when there are no overlapping atoms or all possible atomic interactions are collision free. Collision detection (CD) is an essential problem in robotics, computational geometry, and computer graphics. It is also a major bottleneck in any interactive simulation. A wide range of techniques have been proposed to deal with collision detection such as hierarchical representations, spatial partitioning, analytical methods, and geometric reasoning. The algorithmic design depends on the representation of the model, the query types, and the simulated environment [15].

Bounding volume hierarchies (BVH) are the most popular approach in collision detection to capture self-collision and collisions between objects [21]. The key idea is to use a hierarchical structure to capture the shape of an object at successive levels of detail. The object of interest is enclosed by bounding volumes that can have various shapes such as spheres, axis-aligned bounding boxes (AABBs), and oriented bounding boxes (OBBs) [21]. These bounding volumes become the tree leaves of the hierarchy and subsequent bounding volumes enclose them forming a tree of bounding volumes.

For molecular structures, collision detection is a computationally expensive problem given the many degrees of freedom that a molecule can have. Lotan et al. [14] built a BVH using object-oriented bounding boxes to enclose the molecule. Agarwal et al., [1] used a BVH with the objects being modeled as spheres to detect collisions for deforming and moving necklaces (sequence of balls/beads). Angulo et al. [2], proposed the BioCD algorithm for efficient self-collision and distance computations using axis aligned bounding boxes (AABBs). Redon et al., [19] presented an adaptive dynamic algorithm for articulated bodies built upon "the divide-and-conquer algorithm" (DCA). Morin and Redon in [17] proposed a force-feedback algorithm for adaptive dynamic simulation.

However, most of the above methods do not address the modeling of molecules for real-time rendering or only allow a limited number of degrees of freedom to change. A more generic methodology is required that

- is not limited to the topology of the molecules for self-collisions or collisions between them
- evaluates arbitrary conformations independently of the previous query for effective molecular conformational search
- is adaptive to the molecular structure by exploiting the fact that when limited degrees of freedom change some of the atomic distances remain constant. Incorporating this concept is vital for reducing unnecessary collision detection queries, and hence, reducing the computational time for identifying molecules feasibility.
- identifies molecular feasibility rapidly, efficiently and is evaluated in terms of both time and accuracy that is essential towards the realistic representation of molecular behavior

In our previous work, a new approach called BioGeoFilter (BGF) was presented to model flexible drug-like molecules in real-time [5]. As shown in Fig. 2, a drug-like molecule (also called a *ligand* molecule) is a small molecule that consists of at most 50 atoms. A ligand molecule has the tendency to bind to a larger molecular structure or macromolecule (also called a *receptor* molecule) leading to the identification of pharmaceutical drugs and new molecular arrangements with specific capabilities. Macromolecules on the other hand consist of hundreds or thousands of atoms with hundreds or even thousands of degrees of freedom making their modeling a very computationally intensive task, as shown in Fig. 2. In addition, intricate characteristics of macromolecules such as size, shape and topology need to be considered. Therefore, this work proposes an improved biologically-inspired geometric methodology called enhanced BioGeoFilter (eBGF) to realistically model flexible macromolecules such as proteins. The proposed eBGF approach addresses current literature limitations by providing an effective filtering tool for the identification of feasible macromolecules in order to minimize molecular conformational search and speed collision detection queries. The presented eBGF methodology can facilitate the modeling of flexible macromolecules that will enable the development of an interactive computer-aided design tool for bionanoscale design.

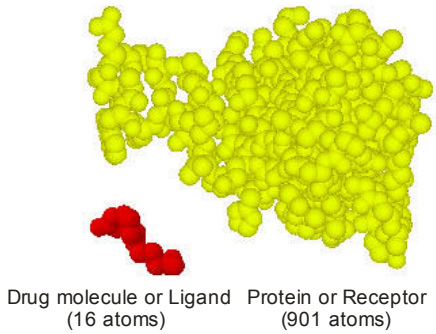


Fig. 2: Receptor and ligand molecules used for drug design.

The remaining of the paper is organized as follows: Section 2 presents the proposed enhanced BioGeoFilter methodology. Then, computer implementation and results are provided in Section 3 followed by the conclusions in Section 4.

2. ENHANCED BIOGEOFILTER METHODOLOGY (eBGF)

The traditional measure of a molecule's feasibility is given by calculating the molecule's internal energy. A feasible molecular conformation indicates a stable molecular state that is described by the minimum possible intramolecular energy. This energy is a complex function composed of different energy factors that depict the interactions between bonded atoms and non-bonded atoms. One of the major energy contributors is the van der Waals (VDW) interaction that models the pair-wise potential over all pairs of non-bonded atoms *i*, *j* as it is shown by the following equations:

$$e_{ij} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}, \tag{2.1}$$

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij} \tag{2.2}$$

Where A_{ij} and B_{ij} are the van der Waals repulsion and attraction parameters, respectively; r_{ij} is the distance between every exclusive non-bonded atom pair *i* and *j*.

The internal energy *E* as shown by Eqn. 2.2 is determined by calculating the pair-wise energy potential e_{ij} between each non-bonded atom pair *i* and *j* within the molecular structure, as shown in Eqn. 2.1. As the number of atoms within the molecule increases, the time to calculate the intramolecular energy for

determining a molecule's feasibility (stability) increases significantly. Since important constraints influencing molecular behavior have geometrical interpretation, this paper presents a new enhanced methodology based upon our previous work called BioGeoFilter, which models the behavior of drug-like molecules [5]. The enhanced BioGeoFilter (eBGF) presented here uses a biologically-inspired geometric approach to effectively identify infeasible conformations for much larger molecules such as proteins. The proposed eBGF approach consists of a hierarchical data structure that comprises two layers: a lower level and an upper level as shown in Fig. 3. At the lower level, the protein is modeled as a highly articulated body with the internal degrees of freedom representing the number of torsion angles. A new approach is also introduced for further simplifying molecular representation and for reducing collision detection search. At the upper level, a bounding volume hierarchy (BVH) depicted as a balanced binary tree is built to identify atoms' self-collisions. The type of bounding volumes selected in this work are spheres as they are simple to test for overlaps and are invariant to rotations.

2.1 eBGF Overview

Fig. 3 shows the overview of the enhanced BioGeoFilter methodology that consists of two layers: the lower and upper hierarchical layers as indicated by the pink colored boxes. At the lower layer of the hierarchy, the eBGF algorithm starts with any molecular conformation. The degrees of freedom (DOF) of the molecular structure are defined to form atom groups following the concept presented in [22]. A further simplification in molecular representation is proposed by splitting the backbone atom cluster into smaller groups of atoms as it is discussed in Section 2.2. At the upper layer of the proposed approach, we build a BVH for the initial molecular conformation as it is described in Section 2.3.1. New random molecular conformations are obtained by arbitrary changing the values for each degree of freedom. For each candidate molecular conformation, the BVH is updated to incorporate the corresponding changes in the DOF as it is presented in Section 2.3.3. A collision detection scheme is then performed to identify the feasibility of each random molecular conformation as it is described in Section 2.3.4. At the end of the eBGF algorithm, the intramolecular energy value for each random conformation is calculated for evaluating the proposed approach as it is discussed in Section 3. The following sections describe in details each hierarchical layer of the proposed eBGF methodology.

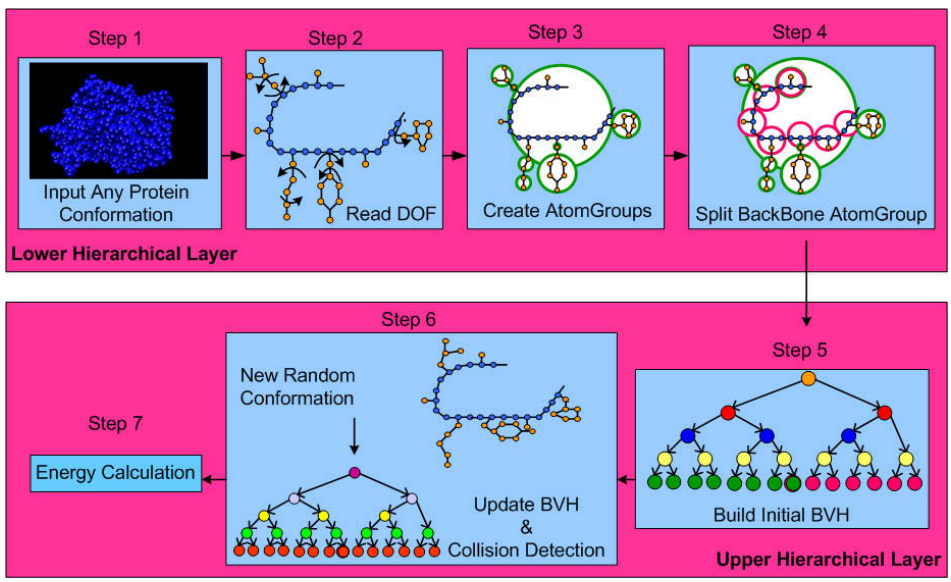


Fig. 3: Overview of the proposed eBGF approach.

2.2 Lower Layer Hierarchy

Proteins are important macromolecules in living organisms that perform thousands of distinct functions. Most proteins are enzymes performing biochemical functions such as bond-making and bond-breaking reactions. Other proteins act as molecular motors or structural components by

performing biophysical functions. Proteins are chains of smaller molecular entities called amino acids. The amino acids consist of a central carbon atom, denoted as C_a, connected to an amino group NH_2, a carboxyl group $COOH$, a single hydrogen atom H, and a side chain R, specific for each amino acid, as it is shown in Fig. 4(a). There are 20 basic amino acids that serve as building blocks of proteins. Each one differs from each other by their side chains, which also determine their chemical characteristics. The amino acids may be linked to each other by the peptide bond (a covalent bond) between an amino group of one amino acid and a carboxyl group of another amino acid releasing a water molecule, as it is shown in Fig. 4(b). These peptide bonds lead to a linear ordering of amino acids forming a polypeptide chain. The backbone of the chain is formed in the sequential pattern schematically shown in Fig. 5. Therefore, any protein can be considered as a polypeptide chain characterized by the amino acid sequence along the chain in order [16].

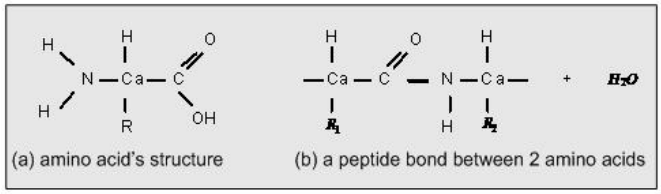


Fig. 4: Graphical representation of amino acids' topology and link procedure through a covalent bond.

Fig. 5: Pattern of a protein's backbone chain.

From a geometric point of view, a protein molecule can be considered as a highly articulated body, where an arbitrarily-selected atom or atom-group acts as the base of the body. Each atom within a protein structure can be considered as a kinematics joint and each bond as a kinematics link. Therefore, a flexible molecule has at least six degrees of freedom (dof): three translational and three rotational. Moreover, some of a protein's bonds have the ability to rotate along its own axes by a torsion angle θ_i that accounts for an additional dof as shown in Figs. 6(a) and 6(b). Bond angles and bond lengths also influence the molecular conformation but are considered constant in this research work as they cause minor changes in the molecular geometry. Thus, a molecular conformation is defined as the changes in the angles of the *k* torsion bonds.

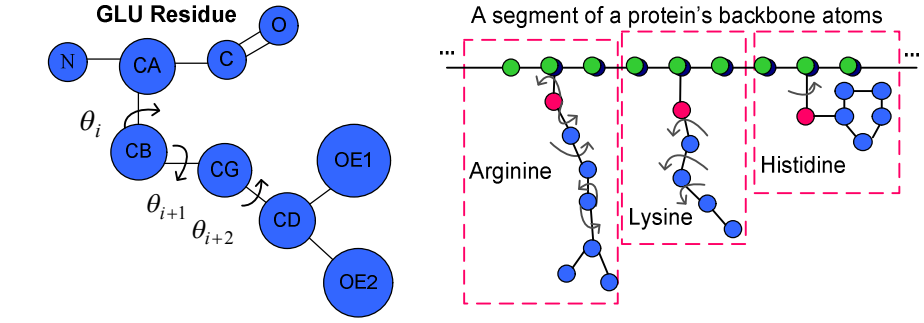


Fig. 6: Graphical representation of the degrees of freedom of a protein.

Torsion changes can occur anywhere within a protein's topology. However, considering random torsions within a protein's backbone can break its structure making it extremely difficult to evaluate whether the generated molecular conformation is chemically feasible. Therefore, in this paper torsions

are assumed only between the central carbon atom of a protein's backbone (*CA*) and a side chain atom (*CB*) or within the side chain atoms as shown in Figs. 6(a) and 6(b). Furthermore, given the increasing complexity by a protein's size, torsions at the end of each side chain are neglected (i.e. the bond between *CD* and *OE1* atoms in Fig. 6(a)) since they do not contribute significantly to the molecular conformation.

Given that the number of possible molecular conformations grows in proportion to the power of the number of torsion bonds, identifying feasible molecular conformations remains the main challenge in molecular design. Therefore, to reduce the computational complexity due to the large number of atoms within a molecular structure, the atoms of a protein are clustered into *AtomGroups* based on the approach proposed by [22]. Based on the location of the torsion bonds, atoms are clustered into AtomGroups. In other words, all the atoms within an AtomGroup are connected by rigid bonds while AtomGroups are connected by torsion bonds as shown in Fig. 7(a).

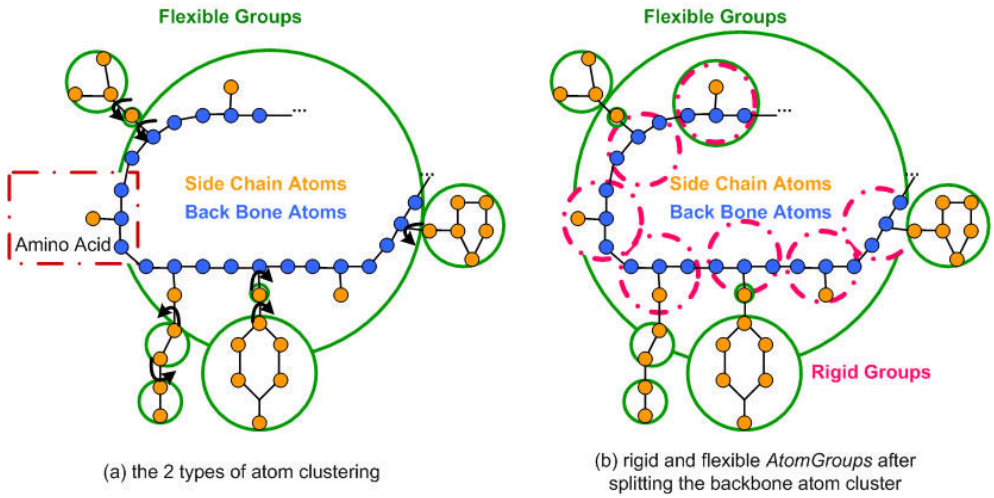


Fig. 7: Graphical representation of the *AtomGroup* concept along with the proposed splitting procedure for a hypothetical protein segment.

Each AtomGroup is assigned a Cartesian coordinate frame and a relationship is generated between all the AtomGroups. Since each AtomGroup contains atoms whose distance will not change when torsion changes occur, the distances between atoms in the same AtomGroups do not need to be checked for collision. Only non-bonded atoms that correspond to different AtomGroups will be checked thus reducing the time to identify geometrically feasible conformations. This significantly reduces the computation time and decreases inaccuracies in calculations for the update of atoms positions during conformational search.

However, the application of the AtomGroup concept in a protein molecule results in the generation of two different sized atom clusters: clusters of side chain atoms and a cluster of backbone atoms as shown in Fig. 7(a). The cluster of backbone atoms contains hundreds of atoms whereas the clusters of side chain atoms contain tens of atoms. This large size difference in the atom clusters increases the time needed to determine if an actual molecular self-collision occurs. In order to address this challenge, this paper proposes to split the backbone atom cluster into smaller AtomGroups based on a threshold defined by the maximum number of atoms allowed within each atom cluster. By splitting the backbone cluster, a flexible AtomGroup (i.e. the green and pink sphere in Fig. 7(b)) is obtained along with a number of rigid AtomGroups (i.e. the six pink/dashed-line spheres shown in Fig. 7(b)). This splitting procedure further simplifies the molecular representation by reducing the collision queries while eliminating the collision searches between the rigid groups. Hence, the collision detection is now performed between similar sized flexible groups of atoms significantly reducing the

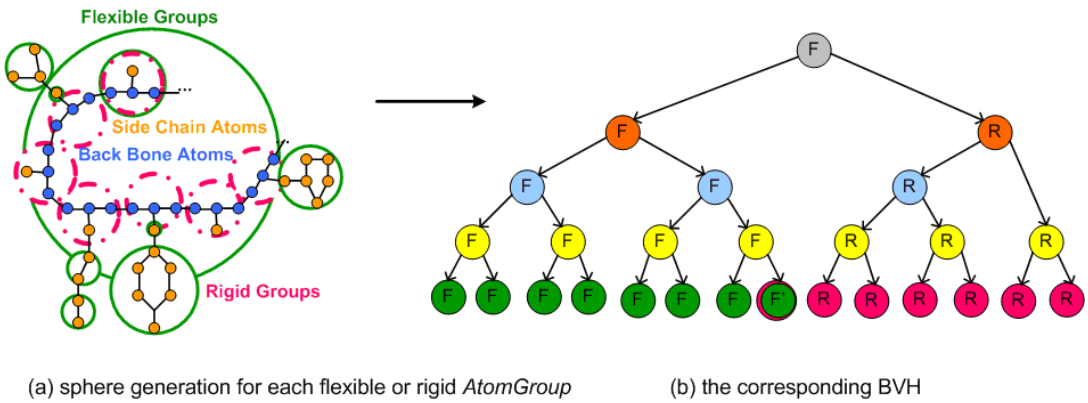
computational time for identifying a molecule's feasibility. In addition, by splitting the backbone cluster into smaller groups of atoms, the computational time for updating the atoms' positions is significantly reduced. This contribution is attained by eliminating the calculation of the relation matrices for the rigid groups of atoms. As shown in Fig. 7(b), the relation matrix for the big green sphere (initial flexible AtomGroup) is the same as the relation matrix of the green and pink sphere (modified flexible AtomGroup) and as the relation matrices of the pink/dashed-line spheres (rigid AtomGroups). Therefore, instead of calculating relation matrices for all the seven new groups of atoms, we just calculate a single relation matrix for the modified flexible group as depicted by the specific protein segment shown in Fig. 7. The clustering of atoms into both rigid and flexible groups will be used to form the upper layer of the hierarchy of the proposed eBGF methodology as it is described in the following section.

2.3 Upper Layer Hierarchy

At the upper level of the proposed hierarchical structure, a bounding volume hierarchy (BVH) depicted as a balanced binary tree is introduced to identify atoms' self-collisions. The BVH is built above the lower layer of the hierarchy to capture the shape of the molecule at successive levels of detail. The type of bounding volume used in this work is a sphere as it simplifies collision detection and is invariant to rotations. New molecular conformations are obtained by randomizing the torsion angle values and then updating the BVH in a bottom up manner. A collision detection algorithm is then performed to search the existence of overlaps between non-bonded atom-pairs within the new molecular conformation. For the same random molecular conformation, the intramolecular energy is also calculated for evaluating the proposed eBGF approach in terms of both time and accuracy as it will be described in Section 3. The following sub-subsections describe in detail the construction and update procedure of the BVH along with the collision detection algorithm.

2.3.1 Constructing the BVH

Once the different AtomGroups (both flexible and rigid) have been defined at the lower layer of the hierarchy, the smallest enclosing sphere that contains all the atoms within each AtomGroup is calculated as in [7]. The spheres (each containing an AtomGroup) are organized into a binary tree-like data structure that will serve to detect molecular self-collisions subject to both chemical and geometric constraints during conformational search as it will be discussed in Section 2.3.4.



(a) sphere generation for each flexible or rigid *AtomGroup* (b) the corresponding BVH

Fig. 8: Schematic representation of the lower hierarchical layer for a hypothetical protein segment and its corresponding BVH.

Fig. 8(a) shows a hypothetical protein segment with its corresponding bounding volume hierarchy shown in Fig. 8(b). At the bottom of the tree are the 14 spheres (called leaf nodes) representing the 14 AtomGroups of the molecule that includes flexible (green colored) and rigid (pink colored or dashed-lines). For each pair of nodes, an intermediate node is created that encloses the two nodes. This process continues in a bottom-up manner until all the spheres result into one single root sphere as

shown by the purple colored sphere in Fig. 8(b), which is the sphere that encloses the whole protein segment.

The BVH is built only once at the beginning of the algorithm allowing a total construction time of $O(N)$ where

$$N = DOF + 1 + rigidGroups = flexibleGroups + rigidGroups$$
$$TotalNumberOfNodes = \begin{cases} 2N & if \quad N \quad odd \\ 2N-1 & o/w \end{cases} \tag{2.3}$$

2.3.2 Randomization

As soon as the pre-selected dof have been defined for the specific protein molecule, a uniform generator is used to create random values for the torsion angles θ_i, where $\theta_i \in [0,2\pi)$. When random torsion changes occur, the new atom positions are updated based on the concept presented in Section 2.2 to obtain a new molecular conformation. For each new random molecular state, the BVH is updated and the new molecular conformation is tested for self-collision.

2.3.3 Updating the BVH

Every time the torsion bonds change, a new molecular conformation is generated. The new atom positions affect the location and radius of the spheres in the hierarchy so they need to be updated accordingly. In this work, the spheres in the BVH are updated in a bottom-up manner and one level at a time. Therefore, the tree nodes are updated from the leaf nodes to their parents until the root node of the tree is reached and updated.

However, as shown in Fig. 8(b), the BVH is formed by both flexible (green colored) and rigid (pink colored) spheres. It can be noticed that the updating of the spheres around the rigid groups (pink colored spheres and their parents) can be neglected since the atom distances within and between the rigid groups remain unchanged. This occurs when the atom cluster of the molecule's backbone has been defined as the root AtomGroup or else the base of the molecule's body. Omitting the update of the rigid nodes (*k*) results in a reduction of the computational time for updating the BVH and for identifying molecular feasibility. This contributes to a total updating time of $O(\frac{N}{k})$ that never exceeds $O(N)$.

2.3.4 Collision Detection

The fundamental concept underneath the proposed collision detection algorithm is the geometric interpretation of the chemical information provided by the van der Waals (VDW) interaction as it has been described in our previous work [5]. The main difference lies in that the collision search presented in this paper is handled independently from the BVH update procedure. In other words, the BVH is updated first and then the tree is traversed down (in a top-bottom mode) to check for possible overlapping atoms.

The geometric constraints used in this work to handle the collision detection queries are depicted by Eqn. 2.4:

$$d_{spheres_{ij}} < \rho_1 (sphereRadius_i + sphereRadius_j)$$
$$d_{atoms_{ij}} < \rho_2 (atomRadius_i + atomRadius_j) \tag{2.4}$$
$$0 < \rho_1, \rho_2 \leq 1$$

Where, $d_{sphere_{i,j}}$ denotes the distance between the sphere objects *i* and *j*; $d_{atoms_{i,j}}$ symbolizes the distance between the non-bonded atoms *i* and *j*; ρ_1 and ρ_2 are constants that control the proposed algorithm's selectivity mechanism.

The first constraint in Eqn. 2.4 embodies a primary filtering while checking for possible collisions between two spherical objects. If this constraint is satisfied, then a possible collision occurs between the sphere objects *i* and *j*. The second constraint in Eqn. 2.4 ensures that an actual self-collision occurs by comparing pair-wise atomic distances. The physical interpretation of the ρ_1 parameter is that it controls the not so tight object fitting that result from the selection of spheres as the type of bounding volumes. The ρ_2 parameter controls the algorithm's selectivity or the number of feasible solutions generated. By decreasing the value of ρ_2 selectivity parameter, the proposed eBGF algorithm accepts more solutions (molecular conformations) as feasible. From a biological point of view, ρ_2 handles the impact that the VDW equilibrium distance has on the results.

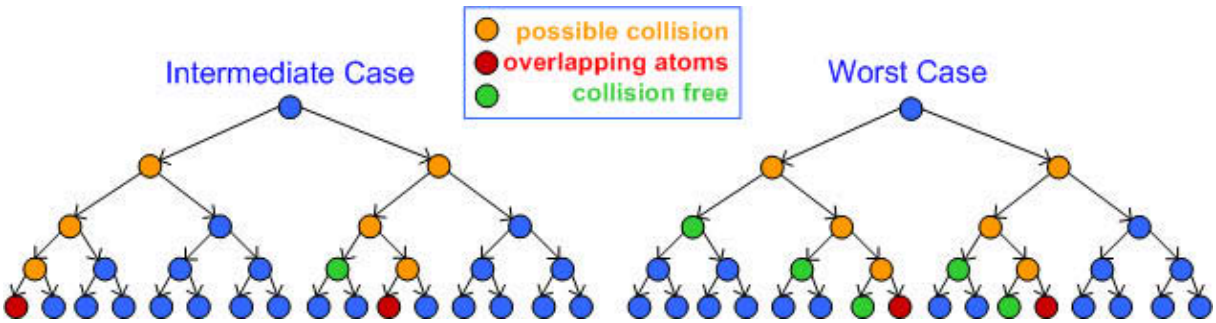


Fig. 9: Graphical representation of the proposed collision detection algorithm.

Fig. 9 shows an intermediate and the worst case scenario of the proposed collision detection scheme for a protein segment. During the tree traversal, each non-constraint pair of nodes is checked for a possible collision using the first constraint in Eqn. 2.4, where the actual self-collision detection is performed between non-bonded atom pairs by using the second constraint in Eqn. 2.4. A constrained node pair embodies any of the following properties:

1. In view of studying self-collision queries, collision detection between the root of the tree against itself should be omitted.
2. The collision search between rigid AtomGroups should be ignored since the atomic distances within and between these groups remain unchanged as it has been analyzed in Section 2.2.
3. Collision queries between bonded neighboring AtomGroups should also be eliminated since the atomic distances between these two groups do not change significantly.
4. Given that the impact of the VDW interaction increases as the pair-wise atomic distances decreases, the collision detection between any atom pair linked by 3 or less chemical bonds should be avoided as in [5]. Therefore, non-bonded atoms are the atoms linked by 4 or more chemical bonds.

Under these assumptions, if the root's child nodes are collision free, then the specific molecular conformation is feasible and is accepted. Otherwise, the tree is traversed down to identify whether any atoms are actually in collision to reject the current molecular conformation. The computation collision detection time has $O(\log\frac{N}{k})$ performance that never exceeds $O(\log N)$, where k is a constant that represents the number of constraint nodes (i.e. rigid AtomGroups) that are neglected in the proposed collision detection scheme. Finally, each random molecular conformation is tested with both the proposed eBGF approach and the traditional energy calculation method using Eqn. 2.2. The main reason for calculating the intramolecular energy in this paper is to evaluate the proposed eBGF methodology in terms of time and accuracy. Further discussion is provided in the following section.

3. COMPUTER IMPLEMENTATION AND RESULTS

The presented method and algorithms have been implemented on a dual 3 GHz CPU workstation using Visual C++ programming language, OpenGL and CGAL libraries [6]. Two different protein molecules

with different number of atoms, residues, and number of degrees of freedom have been tested using the proposed eBGF methodology as shown in Fig. 10. The molecules were obtained from the Protein Data Bank (PDB) [4] with PDB IDs as follows: 1STP and 1DO3 protein molecules. They are displayed using the VMD package [9].

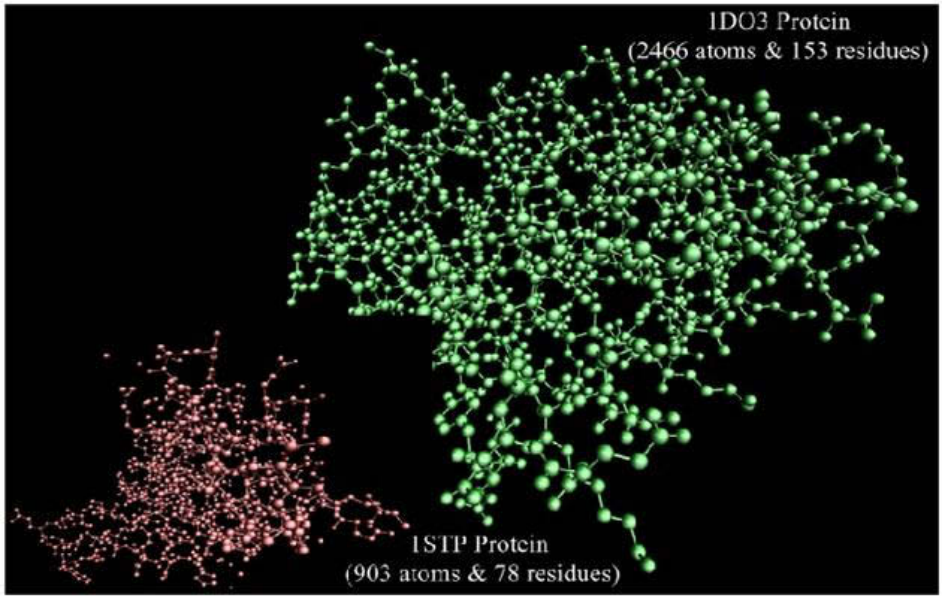


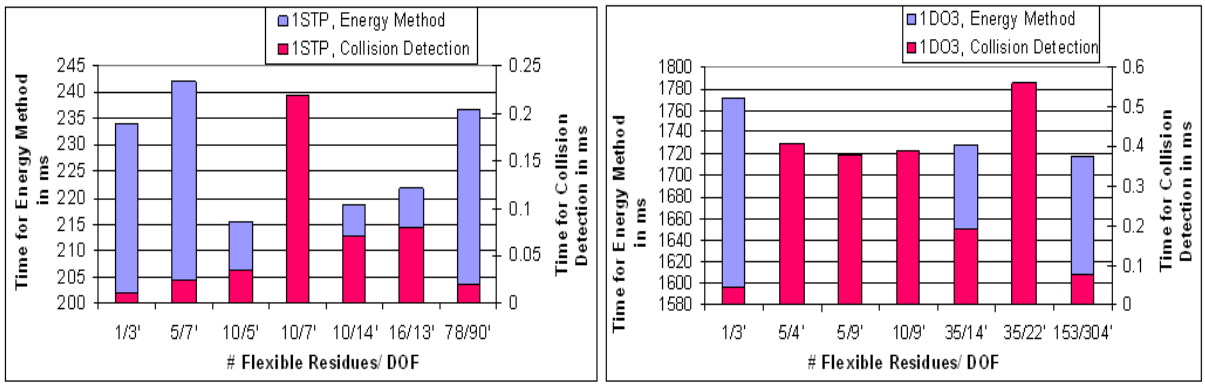
Fig. 10: Two example macromolecules tested in this work displayed using the VMD package.

Molecule	# atoms	Time Energy	Time BVHUpdate	Time Collision	Time Rand	Feasible Energy	TH	Feasible Collision	p1	p2	Free Residues	DOF
		212.42	2.42	0.13	0.3	100/100		100/100	0.4	0.8	1	3
		210.95	1.78	0.02	0.28	94/100		98/100	0.4	0.8	5	7
		218.89	1.99	0.07	0.31	76/100		74/100	0.4	0.8	10	14
		230.71	2.39	0.14	0.33	0/100		0/100	0.4	0.8	16	25
		221.74	1.99	0.079	0.31	0/100		0/100	0.4	0.8	16	13
		224.45	2.34	0.43	0.32	25/100		90/100	0.5	0.7	16	10
		212.85	2.02	0.78	0.29	25/100		25/100	0.6	0.6	16	10
1STP	903	236.78	2.5	0.02	0.38	0/100'	10000	0/100'	0.5	0.9	78	90
		1771.99	6.12	0.044	0.89	10/100		10/100	0.4	0.7	1	3
		1649.68	5.27	0.38	0.78	40/100		75/100	0.5	0.7	5	9
		1647.38	5.45	0.72	0.77	15/100		44/100	0.5	0.6	10	24
		1707.17	5.1	0.24	0.83	0/100		0/100	0.5	0.9	35	36
		1728.51	6	0.19	0.85	0/100		0/100	0.5	0.8	35	14
		1695.27	5.59	0.56	0.81	28/100		29/100	0.5	0.6	35	22
1DO3	2466	1718.19	7.72	0.078	0.94	0/100	100000'	0/100	0.6	0.6	153	304

Tab. 1: Performance analysis of the proposed eBGF algorithm for two example protein molecules.

Tab. 1 shows a representative list of the performance analysis for the proposed eBGF method applied to the two example macromolecules. The molecules have been tested for feasibility after random torsion changes have occurred. The same conformations for both molecules have been examined with both the energy (*TimeEnergy*, *FeasibleEnergy*, and *TH* columns) and eBGF (*TimeBVHUpdate*, *TimeCollision*, *TimeRand*, *FeasibleCollision* columns) approaches and compared in terms of computational time (in milliseconds) and accuracy (percentage of feasible conformations identified). Furthermore, different case scenarios regarding the number, arrangement and the location of the

preselected dof have been tested for assessing their impact on the proposed eBGF methodology (*FreeResidues* and *DOF* columns). Column *FreeResidues* indicates the allowed number of completely flexible residues and column *DOF* represents the total number of dof assumed. For example, in 1DO3 protein section at the bottom of Tab.1: column-pair 35-36 (*FreeResidues-DOF*) indicates that 35 completely flexible residues have been tested for the 1DO3 protein that results in a total of 36 dof; the pair 35-14 indicates that 14 dof were tested only between backbone and side chain atoms; and the pair 35-22 corresponds to torsions only within the side chains of the 35 flexible residues. Further discussion of the impact in molecular behavior by the preselected number of flexible residues and dof is performed below. In addition, different values for the algorithm's selectivity control parameters (ρ_1 , and ρ_2 columns) have been tested and discussed below.



(a) 1STP protein molecule (b) 1DO3 protein molecule

Fig. 11: Comparison of the average collision time by the proposed eBGF vs. the average energy calculation time for different selected sets of preselected flexible-residues/dof.

Fig. 11 compares the performance of the eBGF method against the energy calculation approach in terms of computational time needed to identify molecules' feasibility for the two protein (1STP and 1DO3) examples. As shown in Fig. 11, the eBGF algorithm significantly reduces the computational time needed to identify feasible molecular conformations compared to the energy approach. In fact, the time reduction is so enormous that two different scales were needed to schematically display the two methods in the same graph. The left scale for both figures denotes the time in milliseconds (ms) required by the energy approach to determine the feasibility of a molecular conformation whereas the right scale denotes the computational time (in ms) for the proposed collision detection algorithm to identify molecular feasibility. For both molecules and in all tested sets of preselected flexible-residues/dof, the eBGF requires less than 1ms to output if the tested molecular conformation is feasible. This time reduction is noteworthy as multiple flexible molecules will need to be modeled in real-time simultaneously for the molecular assembly or molecular docking problems.

Similarly, Fig. 12 compares the computational time performance for the two methods (eBGF vs. energy calculation) while considering the scenario that both protein molecules are completely flexible (the total number of residues forming each protein structure assumed to be completely flexible). Analytically, Fig. 12 displays the total computational time (in ms) for the eBGF method (the collision detection time + BVH update time + update atoms' position time) against the energy approach in a logarithmic scale. As it is shown in Fig. 12, the proposed eBGF methodology is significantly faster than the energy calculation approach in identifying feasible molecular conformations. In addition, the eBGF algorithm scales very well as the protein size and problem's complexity increases.

Likewise, Fig. 13 measures the accuracy (percentage of feasible conformations identified) of our proposed method under different considerations regarding the allowed number of flexible residues within each protein structure. For both protein examples, the two methods demonstrate similar

accuracy. In fact, the selectivity of the eBGF algorithm can be adjusted by varying the control parameters (ρ_1 , and ρ_2 presented in Eqn. 2.4).

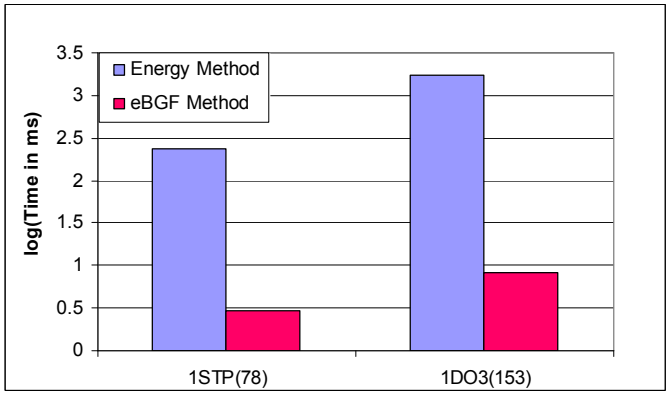
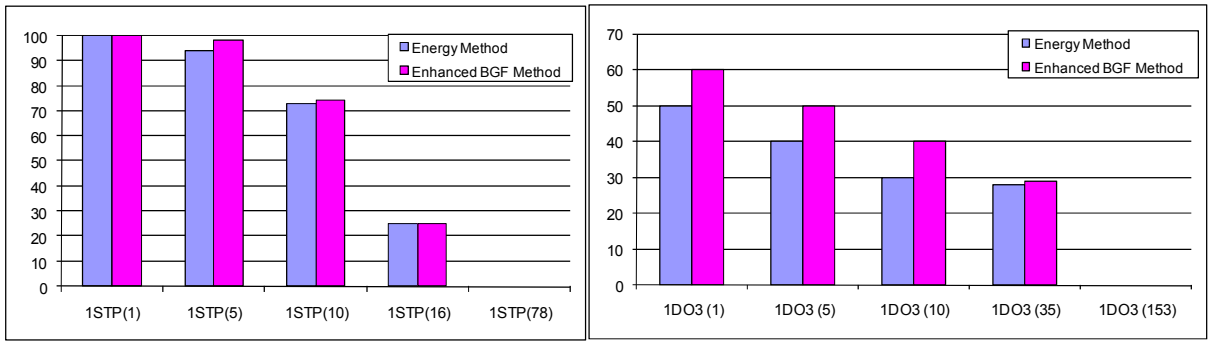


Fig. 12: Average total time comparison between the proposed eBGF algorithm and the energy calculation approach to output feasibility for 1STP and 1DO3 proteins, in a logarithmic scale.



(a) 1STP protein molecule (b) 1DO3 protein molecule

Fig. 13: Schematic demonstration of the accuracy of the proposed eBGF methodology.

In other words, by decreasing the ρ values, the proposed algorithm can accept more molecular conformations as feasible leading to a relaxed filtering. The physical interpretation of ρ_1 selectivity parameter is that it handles the not so tight object fitting resulted by the selection of spheres as the type of bounding volumes whereas the ρ_2 selectivity parameter controls the impact that the VDW equilibrium distance has on the results as it has been analyzed in Section 2.3.4. The main objective of the eBGF approach is to identify infeasible molecular conformations rapidly while not rejecting any feasible ones. Hence, the selection of appropriate ρ values for each molecule depends on the molecule's size along with the desired level of selectivity by the user. It is also worth mentioning here that the molecules' feasibility is traditionally measured using the energy calculation shown in Eqn. 2.2. A conformation might be considered as feasible or not depending on the molecular internal energy value. If a candidate conformation has negative internal energy, then it corresponds to a stable molecular state. However, feasible molecular conformations exist while having positive intramolecular energy. Therefore, a threshold (*TH* column in Tab. 1) has been selected based on the protein's size to define the maximum energy value for which a molecular conformation is considered to be feasible.

Moreover, as it is shown in Tab.1 and Fig. 13, there is a significant dependency among the preselected number of flexible residues and dof considered in each protein molecule and the output (percentage of

feasible molecular conformations) derived by both (eBGF and energy) methods. Computer implementation and results demonstrates that as the number of dof considered increases, the output set of feasible solutions obtained by the energy approach decreases; whereas the output set by the eBGF algorithm can be adjusted as it has been discussed previously. In addition, when many dof are assumed between backbone atom and side chain atoms, the output set of feasible solutions by the energy calculation approach decreases. Therefore, an additional direct search method is essential to identify arbitrarily low energy molecular conformations after they have been filtered by the proposed eBGF methodology.

Tab. 2 demonstrates the worst case scenarios in terms of computational complexity for eBGF and current methods in the literature. The proposed eBGF methodology requires $O(N)$ performance for building and updating the BVH and never exceeds $O(\log N)$ when searching for overlapping atoms. Hence, the eBGF algorithm succeeds to keep the BVH complexity in the lower level ($O(N)$) while significantly reducing collision detection complexity from $O(N)$ to $O(\log N)$.

Methods	Build BVH	Update BVH	Collision Search
ChainTree [Lotan et.al. 2002]	O(N)	O(N)	$O(N^{4/3})$
SpatialAdaptiveHierarchy [Angulo, et.al. 2005]	----	O(NlogN)	O(N)
DeformingNecklaces [Agarwal, et.al. 2004]	O(NlogN)	O(NlogN)	O($N^{4/3}$)
Proposed EnhancedBioGeoFilter (eBGF)	O(N)	O(N)	O(logN)

Tab. 2: Performance analysis of current approaches.

4. CONCLUSIONS

This paper presents an improved biologically-inspired geometric methodology called enhanced BioGeoFilter (eBGF) for modeling the behavior of macromolecules such as proteins. The proposed approach is presented as a rapid filtering tool for the identification of molecules' feasibility (stability). The eBGF algorithm has been tested against the traditional energy calculation approach in terms of computational time and accuracy under different cases. Computer implementation and results demonstrate that the proposed eBGF algorithm significantly decreases the computational time for identifying feasible molecular conformations without sacrificing accuracy. Therefore, the eBGF method facilitates the modeling of flexible macromolecules that will enable the development of an essential interactive computer-aided design tool for bionanotechnology. In addition, the ideas behind the proposed approach can also be applied in areas such as robotics, CAD modeling and virtual surgical simulation where a collision detection scheme is essential.

5. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (NSF) Grant (CMMI-0841451) to the University of South Florida. Their support is greatly appreciated. The authors would also like to thank Dr. Alfredo Cardenas from the Department of Chemistry at USF, Dr. Les Piegl and Roy Soumyaroop from the Department of Computer Science and Engineering at USF for their helpful discussion and suggestions.

6. REFERENCES

[1] Agarwal, P. K.; Guibas, L.; Nguyen, A.; Russell, D.; Zhang, L.: Collision detection for deforming necklaces, Computational Geometry: Theory and Applications, 28(2-3), 2004.

[2] Angulo, V. R.; Cortez, J.; Simeon, T.: BioCD: an efficient algorithm for self-collision and distance computation between highly articulated molecular models, Conference on Robotics: Science and Systems, Boston, MA, 2005.
[3] Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D.: Flexible docking using tabu search and an empirical estimate of binding affinity, Proteins: Structure, Function, and Genetics, 33, 1998, 367-382.
[4] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E.: The Protein Data Bank, Nucleic Acids Research, 28, 2000, 235-242.
[5] Brintaki, A.; Lai-Yuen S. K.: BioGeoFilter: A Tool For Identifying Geometrically Feasible Molecular Conformations In Real-Time For Bionanomanufacturing, Transactions of the North American Manufacturing Research Institution of SME, NAMRC 36, Monterrey, Mexico, 2008, 153-160.
[6] CGAL, http://www.cgal.org, Computational Geometry Algorithms Library.
[7] Fischer, K.; B. Gartner, The Smallest Enclosing Ball of Balls: Combinatorial Structure and Algorithms, SoCG'03, San Diego, California, June 8-10, 2003.
[8] Grayson, P.; Tajkhorshid, E.; Schulten, K.: Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics, Biophysical Journal, 85, 2003, 36-48.
[9] Humphrey, W.; Dalke, A.; Schulten, K.: VMD-Visual Molecular Dynamics, Journal of Molecular graphics, 1999; 14: 33-38.
[10] Lai-Yuen, S. K.; Lee, Y.-S.: Interactive computer-aided design for molecular docking and assembly, Computer-Aided Design and Applications, 3(6), 701-709, 2006.
[11] Lai-Yuen, S. K.; Lee, Y.-S.: Energy-field optimization and haptic-based molecular docking and assembly search system for computer-aided molecular design (CAMD), Proceedings of the 14th Symposium Haptic Interfaces for Virtual Environment and Teleoperator Systems, IEEE Virtual Reality Conference, Alexandria, VA, 25-29, 2006.
[12] Lai-Yuen, S. K.; Lee, Y.-S.: Computer-aided design and simulation for nano-scale assembly, Transactions of the North American Manufacturing Research Institution of SME, 34, 2006, 357-364.
[13] Lee, Y-G.; Lyons, K. W.: (2004). Smoothing Haptic Interaction using Molecular Force Calculations, Computer-Aided Design, 36 (1), 2004, 75-90.
[14] Lotan, I.; Schwarzer, F.; Halperin, D.; Latombe, J. C.: Efficient Maintenance and self-collision testing for kinematic chains, SoCG, Barcelona, Spain, 2002.
[15] Lin, M. C.: Collision detection between geometric models: a survey, Proceedings of IMA Conference on Mathematics of Surfaces, 1998.
[16] Molecular Biology, http://www.web-books.com/MoBio/, Web Book.
[17] Morin, S.; Redon, S.: A Force-Feedback Algorithm for Adaptive Articulated-Body Dynamics Simulation, IEEE International Conference on Robotics and Automation, Rome, Italy, 2007.
[18] Nagata, H.; Mizushima, H.; Tanaka, H.: Concept and prototype of protein-ligand docking simulator with force feedback technology, Bioinformatics, 18 (1), 2002, 140-146.
[19] Redon, S.; Galoppo, N.; Lin, M. C.: Adaptive dynamics of articulated bodies, ACM Transactions on Graphics, 24(3), 2005.
[20] Stone, J.; Gullingsrud, J.; Grayson, P.; Schulten, K.: A system for interactive molecular dynamics simulation, ACM Symposium on Interactive 3D Graphics, Ed. By J. F. Hughes and C. H. Séquin, ACM SIGGRAPH, New York, NY, 2001, 191-194.
[21] Teschner, M.; Kimmerle, S.; Heidelberger, B.; Zachmann, G.; Raghupathi, L.;, Fuhrmann, A.; Cani, M.-P.; Faure, F.; Magnenat-Thalmann, N.; Strasser, W.; Volino, P.: Collision Detection for Deformable Objects, Computer Graphics forum, 24(1), 2005, 61-81.
[22] Zhang, M.; Kavraki, L. E.: A new method for fast and accurate derivation of molecular conformations, Journal of Chemical Information and Computing Sciences, 42, 2004, 64-70.