JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

SQUEEZE-E: The Optimal Solution for Molecular Simulations with Periodic Boundary Conditions

Tsjerk A. Wassenaar,*[,†,‡] Sjoerd de Vries,[‡] Alexandre M. J. J. Bonvin,[‡] and Henk Bekker[§]

[†]Molecular Dynamics Group, Groningen Institute for Biotechnology and Biomolecular Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands

[‡]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands

[§]Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, POB 800, 9700 AV, Groningen, The Netherlands

ABSTRACT: In molecular simulations of macromolecules, it is desirable to limit the amount of solvent in the system to avoid spending computational resources on uninteresting solvent−solvent interactions. As a consequence, periodic boundary conditions are commonly used, with a simulation box chosen as small as possible, for a given minimal distance between images. Here, we describe how such a simulation cell can be set up for ensembles, taking into account *a priori* available or estimable information regarding conformational flexibility. Doing so ensures that any conformation present in the input ensemble will satisfy the distance criterion during the simulation. This helps avoid periodicity artifacts due to conformational changes. The method introduces three new approaches in computational geometry: (1) The first is the derivation of an optimal packing of ensembles, for which the mathematical framework is described. (2) A new method for approximating the α-hull and the contact body for single bodies and ensembles is presented, which is orders of magnitude faster than existing routines, allowing the calculation of packings of large ensembles and/or large bodies. 3. A routine is described for searching a combination of three vectors on a discretized contact body forming a reduced base for a lattice with minimal cell volume. The new algorithms reduce the time required to calculate packings of single bodies from minutes or hours to seconds. The use and efficacy of the method is demonstrated for ensembles obtained from NMR, MD simulations, and elastic network modeling. An implementation of the method has been made available online at http://haddock.chem.uu.nl/services/SQUEEZE/ and has been made available as an option for running simulations through the weNMR GRID MD server at http://haddock.science.uu.nl/enmr/services/GROMACS/main.php.

INTRODUCTION

In order to avoid surface effects in molecular simulations, periodic boundary conditions (PBCs) are commonly used. PBCs allow mimicking an infinite system, taking a single simulation system or unit cell and surrounding it with replica cells in a space-filling way. With respect to the computational cost, it is desirable to choose an optimal unit cell. In this regard, optimal is usually interpreted as minimal in total volume, under the condition that a prescribed minimal distance exists between any two periodic images. Alternatively, optimal could also be considered as having a maximal separation between periodic images, given a certain volume of the system. This would in effect minimize direct self-interactions as well as indirect self-interactions mediated by water ordering and allow larger conformational changes.

For a solute which is approximately spherical or which can rotate freely, the optimal simulation cell is based on a rhombic dodecahedron: the packing of the circumscribed sphere. However, when rotational degrees of freedom are removed, it is also possible to use the molecular geometry of a (nonspherical) solute to determine an optimal simulation cell. Using the so-called *near-densest lattice-packing* (NDLP)[1] method, almost all redundant solvent can be removed, where redundancy is defined as being further away than half the prescribed minimal distance from any of the periodic images of the solute. Such a simulation cell offers two advantages. First, as originally reported, the volume of the simulation system can be greatly reduced, with an average of 55% for a random series of proteins, as compared to the corresponding rhombic dodecahedron box set up with the same minimal distance, resulting in a speed-up factor for the simulations of about 2.5. But next to reducing the simulation volume, the NDLP method also inherently yields a system with a regular distribution of solvent in between periodic images. Irregularities of the solvent distribution have been suggested to account for the differences observed in simulations performed in different box types,[2] and this would be alleviated by choosing PBCs having a regularized distribution, such as the rhombic dodecahedron and the NDLP.

Removal of the rotational degrees of freedom requires the implementation of a constraint algorithm. Mathematical proof has been given that rotational degrees of freedom can be removed without affecting the statistical mechanical ensemble,[3] and it has been demonstrated that the use of an NDLP unit cell in conjunction with rotational constraints does not lead to detectable changes relative to simulations performed in a rhombic dodecahedron.[2]

Special Issue: Wilfred F. van Gunsteren Festschrift

Received: January 30, 2012
Published: May 28, 2012

dx.doi.org/10.1021/ct3000662 | *J. Chem. Theory Comput.* 2012, 8, 3618−3627

Unfortunately, the NDLP method is slow and memory demanding. Especially for large molecules, the time to compute the simulation cell is significant, even though it is not comparable to the speed-up of the simulations. In addition, the NDLP method determines a minimal volume unit cell based on a single structure, represented as a set of points, and any significant deviation from that structure during the simulation is likely to cause violations of the minimal distance required between periodic images. Although this is a concern with all simulation cells, it is more profound with a tight-fitting NDLP unit cell.

A naïve approach to solve this problem is to take an estimate of the ensemble, e.g., as provided by NMR or obtained from elastic network modeling, and use the combined atom positions to determine the NDLP. However, the large resultant point set in practice poses problems and will not actually yield an optimal packing with respect to the ensemble. To illustrate this last point, it should be borne in mind that, when using PBC, any conformation observed in one unit cell is by definition surrounded by identical copies. This means that for two possible conformations A and B one does not need to consider the possibility that B occurs in a unit cell neighboring one containing A. The mathematically optimal simulation cell thus is the one with the minimal volume which holds that for every conformation at least the prescribed minimal distance exists between neighboring periodic images. In contrast, determining the NDLP from a combined point set would not only consider the cases of identical neighbors but would also encompass all possible combinations of different neighbors.

Mathematically, the problem can be summarized as follows: Given a set of shapes, determine the set of vectors with a minimal determinant that form a base for a lattice, such that none of the shapes has overlaps with its periodic images. The solution to this computational geometrical problem and the implementation of the algorithms form the main topic of this work, presented within the context of its application: determining optimal volume cells for increased efficiency of molecular simulations with minimal overhead.

As is explained in detail below, the solution requires the calculation of a surface of contact for each shape. This commonly involves the calculation of an α-hull by Delauney triangulation, which makes the procedure slow. To reduce the computational cost, we here introduce a new approach to efficiently determine a contact body. This approach involves scaling the vertices of a preset triangulated surface to the surface of contact.

The article is structured as follows: First, the mathematics involved in constructing a contact-body for an ensemble is given. Then, the practical implementation is described, involving the (fast) generation of an approximate α-hull with a predefined number of boundary points, and the search algorithm for determining the base vectors of the optimal lattice. Finally, the efficacy of the method is demonstrated, and examples are given for setting up a simulation unit cell for an (estimated) ensemble of structures.

For simplicity, most of the figures are representations in two dimensions. The principles described in the text and depicted in these illustrations apply equally well to three dimensions.

MATHEMATICAL BACKGROUND

Construction of the Contact Body. For any two bodies A and B with a fixed relative orientation (i.e., angular position), it is possible to define a surface at which B can be placed relative to A (by translation) such that the two bodies touch, but do not overlap (Figure 1). This surface is called the contact surface and is a central concept in setting up a near-densest lattice packing.[1]

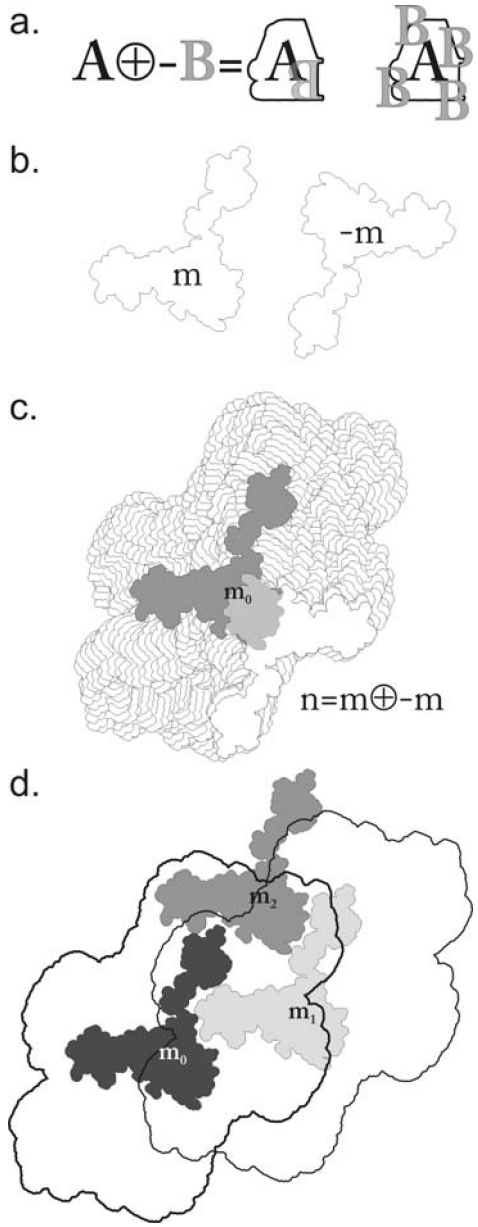

Figure 1. Contact surfaces. (a) For two bodies A and B, the contact surface follows from the outline of the shape obtained by tracing the inverse of B over the outline of A or vice versa. (b) A molecular shape $\mathbf{m}$ and its inverse $-\mathbf{m}$. (c) The contact body $\mathbf{n}$ of $\mathbf{m}$ is obtained by tracing $-\mathbf{m}$ over the outline of $\mathbf{m}$. (d) A translated copy of $\mathbf{m}$ placed on the contact surface will exactly touch but not overlap with the original. Furthermore, a second translated copy placed at the intersection of the contact surfaces of the original body and the first translated copy will touch both without overlapping with either of them.

The contact surface can be obtained as follows: Consider A and B to be two sets of points and let $-B$ denote the set of negated points of B, i.e., $-B \equiv \{-\mathbf{x}: \mathbf{x} \in B\}$. Then, if for each combination of points from A and $-B$ the sum is taken, the set formed by these sums forms the contact body N:

dx.doi.org/10.1021/ct3000662 | J. Chem. Theory Comput. 2012, 8, 3618−3627

$$N \equiv A \oplus -B = \{\mathbf{a} + \mathbf{b}: \mathbf{a} \in \mathbf{A}, \mathbf{b} \in -B\} \quad (1)$$

This operation, denoted by the symbol $\oplus$, is called the Minkowski sum. If A and B are identical, the contact body is symmetric and centered at the origin.

Now, let m denote the solute for which an optimal packing should be defined, given a minimal distance d_m between periodic images. If m is first dilated with half the distance d_m, forming a shape M, it can be easily verified that the surface of the contact body

$$N = M \oplus -M \quad (2)$$

comprises all points over which m can be translated such that the translated body m_1 is separated by exactly a distance d_m from the original (Figure 2).

The dilation itself can be written as the Minkowski sum of m and a sphere R with radius equal to $d_m/2$ (Figure 2A), allowing a rewrite of eq 2

$$N = (m \oplus R_{d_m/2}) \oplus -(m \oplus R_{d_m/2}) = m \oplus -m \oplus R_{d_m} \quad (3)$$

given that the Minkowski sum is commutative and R is symmetric. From eq 3, it follows that it does not matter whether m is dilated first or the dilation is performed on the contact body $n = m \oplus -m$. This is an important note for generating a contact body for an ensemble of structures, as it allows avoiding unnecessary computational steps.

The Contact Body for an Ensemble. To derive the contact body for an ensemble of structures, consider a molecule with p atoms in k configurations (see Figure 3 for $k = 2$). Let $\mathbf{r}_{ij}$ denote the position of atom i in configuration j, and let m_j be the set of p points in 3D space given by

$$m_j \equiv \cup_{i=1}^{p} \mathbf{r}_{ij} \quad (4)$$

such that m_j represents configuration j of the molecule. For determining the contact body, and subsequently the NDLP, of the ensemble, a single point set has to be generated from the different configurations. A naive approach would be simply to take the union of all k configurations:

$$M \equiv \cup_{j=1}^{k} m_j \quad (5)$$

This procedure yields a set of pk points in total. Note that for a given point in M it cannot be determined from which of the k configurations it originates. Accordingly, the contact body N' derived from M using

$$N' \equiv M \oplus -M \quad (6)$$

consists of p^2k^2 points in total. But, as explained before, the combination of points from different configurations in general leads to a nonoptimal packing. To avoid this, instead of calculating a single contact body from the set of points combined, a set of k contact bodies N_j is calculated first, according to

$$N_j \equiv m_j \oplus -m_j \quad (7)$$

Each of these contact bodies consists of p^2 points. Combining them into a single contact body N using

$$N \equiv \cup_{j=1}^{k} N_j \quad (8)$$

yields a points set of size p^2k. This set consists only of points resulting from the combination of pairs of atom positions from identical configurations, according to the requirements for obtaining an optimal packing.

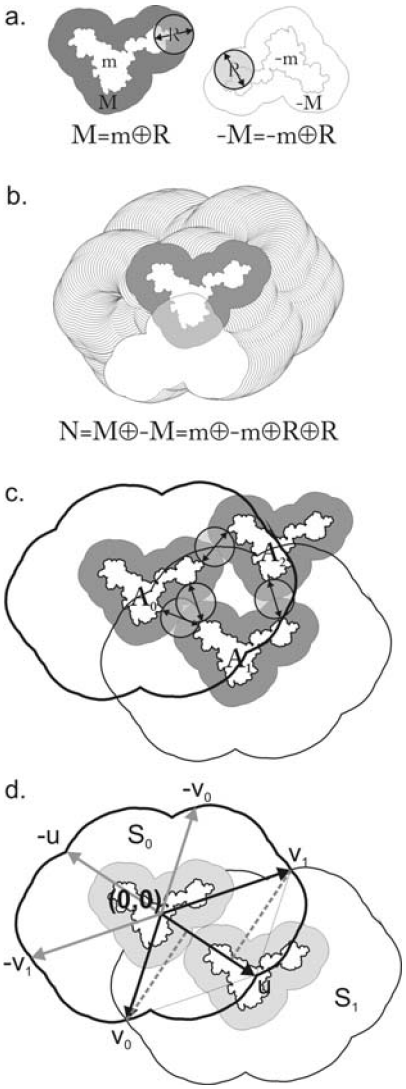

Figure 2. Surface of specified distance and the derived lattice. (a) Dilated body **M**, obtained by taking the Minkowski sum of **m** and a sphere **R**, and the inverse of the dilated body, −**M**. (b) Contact body **N** obtained by tracing −**M** over the outline of **M**. (c) Any translated copy of **m** placed on the surface of **N** will be separated by a distance equal to the diameter of **R**. Furthermore, a second translated copy placed at the intersection of the contact surfaces associated with the original and the first translated copy will be separated from both of these by the distance specified. (d) The points of intersection $\mathbf{v}_1$ and $\mathbf{v}_2$ of a contact body and a translated contact body placed at the surface point **u** form a triple such that $\mathbf{u} = \mathbf{v}_1 + \mathbf{v}_2$. Any pair of these three vectors can be used as a basis for a lattice satisfying the minimal distance criterion.

To introduce dilation, recall from the previous section that the contact body for a point set dilated by a distance $d_m/2$ equals dilating the contact body for the original point set by a distance d_m. Therefore, any dilation need only be performed once, after obtaining the contact body for the ensemble of structures m_j:

$$N^* \equiv R_{d_m} \oplus \cup_{j=1}^{k} N_j \quad (9)$$

From the foregoing it can be verified that the resulting contact body N^* has the property that for the set of translated

dx.doi.org/10.1021/ct3000662 | J. Chem. Theory Comput. 2012, 8, 3618−3627

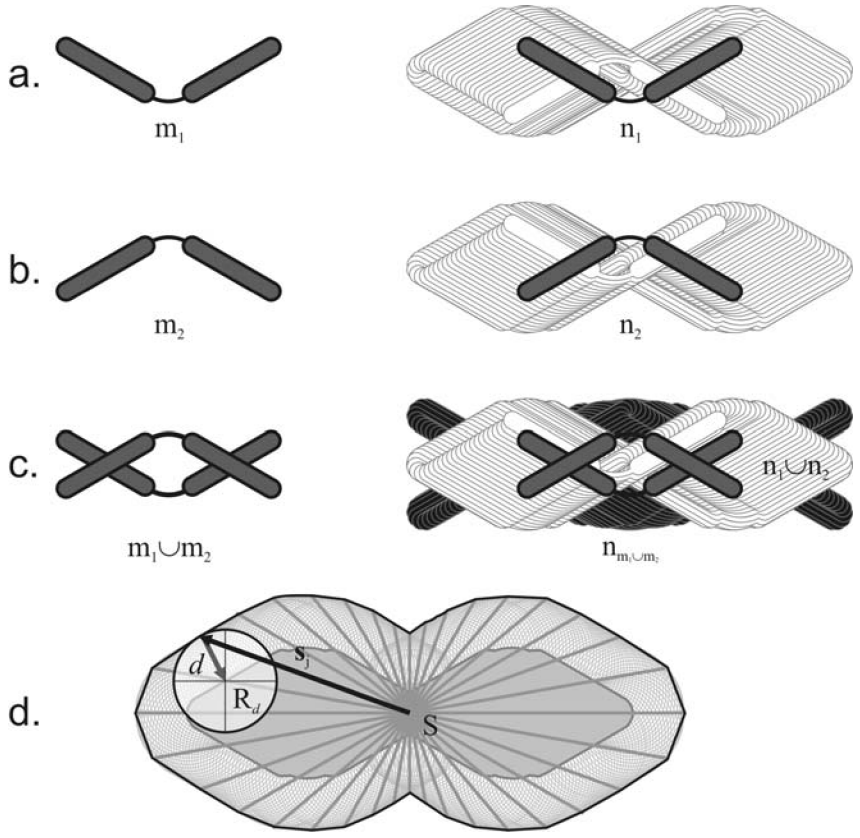

Figure 3. Surface of specified distance for an ensemble of shapes. (a) A shape $\mathbf{m}_1$ and the corresponding contact body $\mathbf{n}_1$. (b) A shape $\mathbf{m}_2$ and the corresponding contact body $\mathbf{n}_2$. (c) The union of $\mathbf{m}_1$ and $\mathbf{m}_2$, $\mathbf{m}_1 \cup \mathbf{m}_2$, and the corresponding contact body $\mathbf{n}_{\mathbf{m1}\cup\mathbf{m2}}$, encompassing both contact bodies $\mathbf{n}_1$ and $\mathbf{n}_2$ (in gray), as well as additional regions (in black). As the black areas arise from the combination of $\mathbf{m}_1$ and $\mathbf{m}_2$, they are redundant and should be excluded when determining an optimal packing for an ensemble of shapes. (d) The contact body $\mathbf{n}_1 \cup \mathbf{n}_2$, and the vector approximation to the surface of minimal distance. For each of a series of vectors $\mathbf{s}_j$ starting from the origin, the end point is determined as the maximum point of intersection of the vector with a sphere placed at the surface of the contact body $\mathbf{n}_1 \cup \mathbf{n}_2$.

bodies m_j placed at its surface there is at least one that is separated by exactly a distance d_m from its original placed at the origin.

Vector Approximation of the Contact Surface. The surface of the contact body can be considered as the set of vectors allowed for translation of m_j. Since the contact body is symmetric and located at the origin, and considering that crevices or holes should in general not be considered for placement of a neighbor, a simple, effective approach can be introduced to approximate the contact surface:

Let $\mathbf{s}$ be a vector starting at the origin. Then, let the length of $\mathbf{s}$ be set such that the end point lies on the contact surface of N. As is illustrated in Figure 3d, this is equal to stating that the length of $\mathbf{s}$ is the maximum length for which the distance between the end point and any point from $n = m \oplus -m$ is equal to the distance d_m. The (triangulated) contact body can thus be approximated by taking a set of vectors, forming a triangulated polyhedron centered at the origin, and scaling each of these accordingly.

From eq 9, it follows that for an ensemble of structures the correct contact surface is obtained by determining the maximum length for each vector over all shapes m_i. Consequently, the approach presented here allows for a determination of the contact body for an ensemble without explicitly calculating a contact body for any structure and avoiding the need to perform calculations on a combined point set.

Searching an Optimal Lattice Base from a Set of Vectors Forming a Contact Surface. Having a representation of the contact surface, the next objective is searching for three vectors forming a basis for a near-optimal lattice, with minimal volume and no overlaps. The standard procedure for this is to iterate over the vertices of the triangulated surface, calculating the line of intersection of the surface placed at the origin and a surface placed at the vertex. Then, for each node on the intersection line, a second copy of the contact surface is placed, and the limited set of points where all three bodies intersect is determined. Each set of three vectors thus obtained defines a lattice in which all direct neighbors touch but do not overlap. From these combinations, the one having a minimal determinant is sought, which corresponds to the optimal packing.

In this section, we present a method for determining the optimal lattice vectors directly from the set of vertices comprising the contact surface, without calculating intersections. First consider that, in the case of two dimensions, the intersection points $\mathbf{v}_1$ and $\mathbf{v}_2$ of the symmetric contact line S_0 placed at the origin, and its copy S_1 placed at a vertex $\mathbf{u}$ on S_0 are the points on S_0 for which $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{u}$ (Figure 2d). Any two of these points can then be used to define the lattice in the plane for that configuration, but only one of these usually corresponds to a reduced lattice, in which the relative projection of one vector on the other is less than or equal to 0.5. We can thus determine a set of valid, reduced 2D lattices by

dx.doi.org/10.1021/ct3000662 | J. Chem. Theory Comput. 2012, 8, 3618−3627

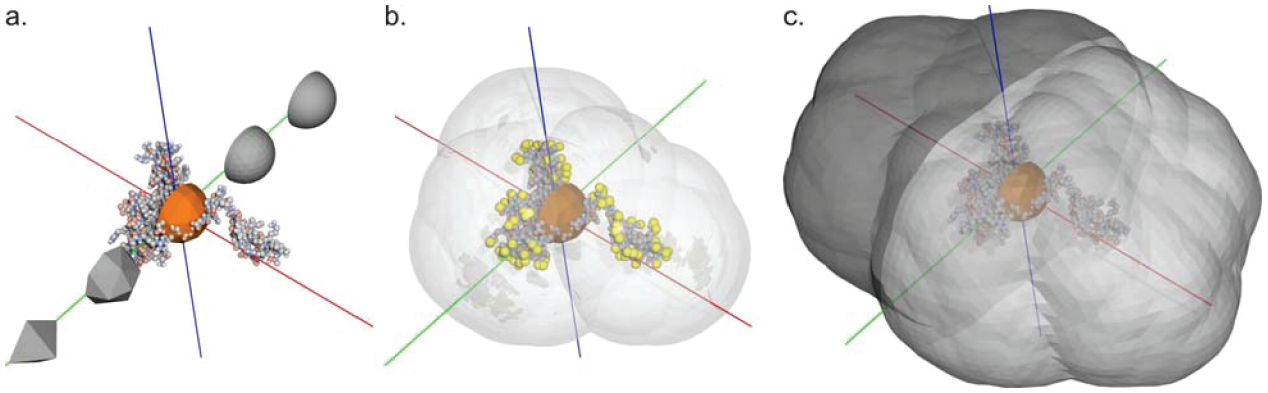

Figure 4. Generation of a three-dimensional contact body. (a) Generation of half sphere approximations starting from a half octahedron, with zero to four triangle division cycles. The half sphere approximation obtained by two triangle division cycles is shown in orange, positioned at the center of a protein (PDB ID: 1A32[5]). (b) Selection of boundary atoms, retaining only those atoms for which at least one of the vectors of the half sphere approximation has its exit point on the sphere with a specified radius, placed at the coordinates of the atom. The boundary atoms are highlighted in yellow and displayed with larger radii. (c) The contact body obtained by scaling the vertices of the half sphere approximation. Only the front half, shown transparent, needs to be determined, as the other half, shown in a darker shade, follows by reflection in the origin.

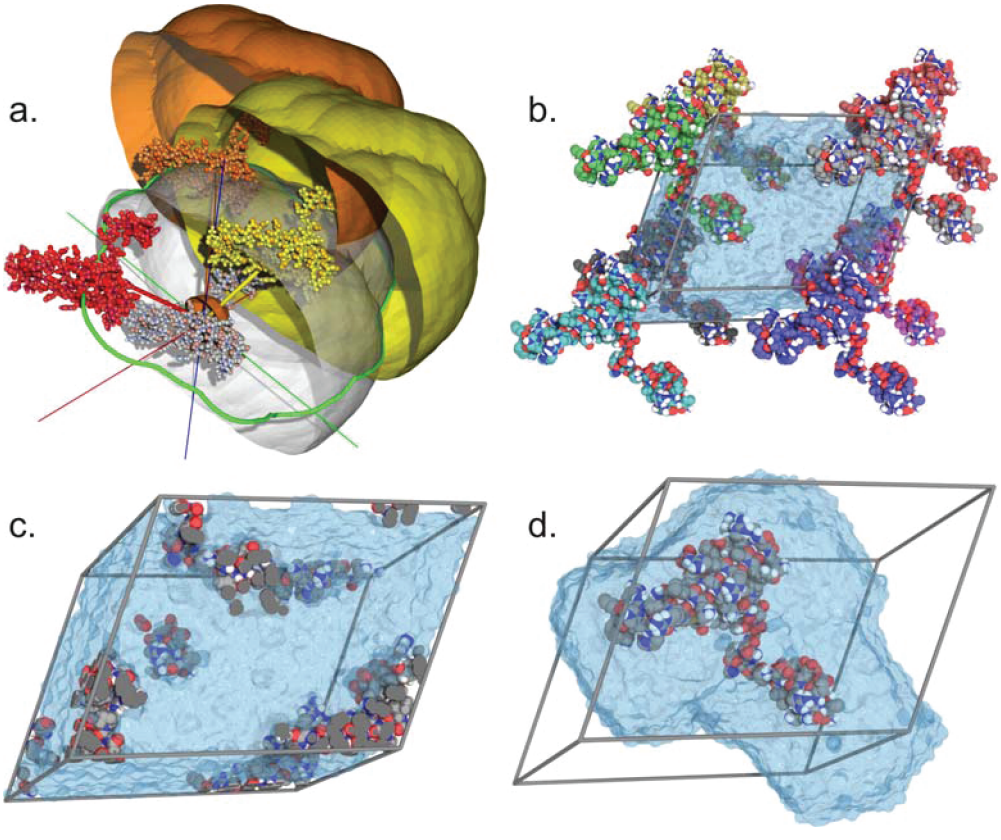

Figure 5. Determining the NDLP and deriving a molecular shaped box. (a) Configuration of four instances of a protein, such that (i) the first translated copy (yellow) lies on the contact body of the original (white), placed at the origin, (ii) the second translated copy (orange) lies on the line of intersection (green) of the contact bodies of the original and the first translated copy, and (iii) the third translated copy (red) is placed at a point of intersection of the contact bodies of the two foregoing translated copies and of the original. (b) The configuration of four instances corresponds to a lattice arrangement, from which a triclinic unit cell can be inferred. (c) If a triclinic unit cell is carved out from the lattice arrangement, the original body may appear broken, and in some cases it will be impossible to translate the system such that the body will entirely fit within the triclinic cell. (d) By carving out the unit cell as the Voronoi region of the protein in the lattice, the "molecular shaped box" can be shown, exemplifying that the body is both whole and entirely surrounded by solvent.

searching proper combinations of vectors comprising S_0, for which either the sum or the difference vector also lies on S_0.

In three dimensions, the aim is finding three independent vectors **u**, **v**, and **w**, together forming a basis for the optimal lattice. Yet in a three-dimensional lattice, each combination (**u**,**v**), (**u**,**w**), and (**v**,**w**) itself defines a two-dimensional lattice. Thus, for obtaining a lattice in three dimensions, we only need to search the combinations of two vectors, such that for the pair (**u**,**v**) a third vector **w** is sought, for which (**u**,**w**) and (**v**,**w**) are also valid lattices. Subsequently, the validity of the lattice

defined by such a combination needs to be assessed by checking that none of the linear combinations of **u**, **v**, and **w** result in a point lying in the interior of S_0. From the set of valid triples of vectors, the one having a minimal determinant corresponds to the near optimal lattice packing.

IMPLEMENTATION

The original implementation[1] for calculating the contact body involved the initial construction of the α-hull[4] of $M = m \oplus R_{d_m/2}$, taking the Minkowski sum of that shape and its inverse, and then calculating the α-hull of the resulting point set to yield the contact body. These operations are costly, even when a grid-based reduction of points is used first, as the α-hull algorithm[4] is quadratic in the number of points. Then, considering that for an ensemble the filtering and determining of the outer hull twice has to be performed for each element, calculating the contact body in this way becomes unacceptably costly.

To increase the efficiency of the procedure, it was desirable to relieve the dependency on the α-hull algorithm. To achieve this, a different approach was taken, involving the vector approximation of the contact surface. In brief, the procedure consists of the following steps (Figure 4): (1) *generation of a triangulated sphere*, (2) *filtering points from* m, and (3) *scaling the vertices to the contact surface*.

For an ensemble of shapes, steps 2 and 3 are repeated for each shape, and the maximum length of each vector is stored, according to the description in the previous section. The different steps are explained in more detail below. Note that determining an approximate α-hull for a point set using this procedure scales linearly with the number of points.

1. *Generation of a triangulated sphere.* The contact body is approximated by scaling the vectors of a triangulated sphere. In fact, because of the symmetry of the contact body, it is sufficient to start with half a sphere. This half-sphere S is obtained from successive subdivision of the triangles forming an octahedron, using simple linear interpolation. Although a better approximation to a sphere can be made,[6] this description suffices for the purpose.
2. *Filtering boundary points of m.* With both m and S placed at the origin, each vertex of S is scaled from the origin to the maximal length for which the vertex is a distance d_m away from any point from m or $-m$. The points of m involved, i.e., the boundary points m^*, are stored, while the other points are discarded.
3. *Calculating the contact body.* Having extracted the boundary points, each vertex from S is scaled from the origin to the maximal length for which the vertex is a distance d_m away from the difference vector of any combination of two points from m^*.

The resulting triangulated contact surface is searched for the three translation vectors corresponding to the near-densest lattice packing, which form the basis for the optimal simulation cell (Figure 5). The whole algorithm can be represented by pseudocode given below. Note that in this summary, the triangulated sphere S is in fact a half-sphere, as the contact body is symmetric. This is also reflected in lines 8 and 9, where K−J and J−K correspond to the symmetric pair of combinations of two points, of which one will lie in the hemisphere corresponding to S. At the end of the routine, the lengths of the vectors comprising S will be set such that S will approximate half of the contact body, with the other half following from the symmetry. As S is already triangulated, it can be directly used for determining the NDLP, according to the three-dimensional search routine developed previously.[1]

```
 1 GENERATE Triangulated Sphere Approximation S
 2 FOR each structure M in ensemble DO
 3  FIT structure to reference
 4  FILTER surface points of structure M
 5  FOR each vector I in S DO
 6       FOR each surface point J of M DO
 7            FOR each surface point K of M DO
 8                 LET R1 be a sphere centered at J-K
 9                 LET R2 be a sphere centered at K-J
10                 IF I has exit point P on R1 or R2 THEN
11                      IF length P > length I THEN
12                           SET I equal to P
13                      END IF
14                 END IF
15            END LOOP // K loop
16       END LOOP // J loop
17   END LOOP // I loop
18 END LOOP // M loop
19 DETERMINE Near Densest Lattice Packing from S
```

RESULTS

Comparison between the Original and Revised NDLP Methods. The results obtained using the new implementation are summarized in Figure 6. Figure 6a shows a comparison of unit cell volumes obtained for the proteins in the PDB in a cubic, a rhombic dodecahedron, or an NDLP unit cell, using a distance between periodic images of 2.5 nm. It is clear that the NDLP unit cell yields the smallest volume at all times. In addition, the results for the rhombic dodecahedron show a large spread in resulting system size for a given solute size, whereas this spread is much smaller for NDLP system sizes. This can be explained from the diversity in solute shapes. A rhombic dodecahedron always relates to the largest diameter, whereas the NDLP adapts the cell to the solute geometry. Note that, for clarity, the y-axis scale is truncated to a volume of 10^4 nm^3.

The decrease in size is further exemplified in Figure 6b, which shows the volume of the NDLP unit cell relative to the corresponding rhombic dodecahedron as a function of the number of atoms in the structure, given on a logarithmic scale to account for the uneven distribution of protein sizes in the PDB. This figure gives an indication of the relative computational cost of simulations set up in either type of box. The horizontal lines indicate the deciles. The top line shows that more than 90% of the structures yield a 40% decrease or more in volume when using an NDLP box instead of a rhombic dodecahedron. This corresponds to a speed-up factor of more than 1.6. In more than half of the cases, the decrease is greater than 50%, more than doubling the simulation speed. It is noted that this increase in efficiency in explicit solvent simulations is

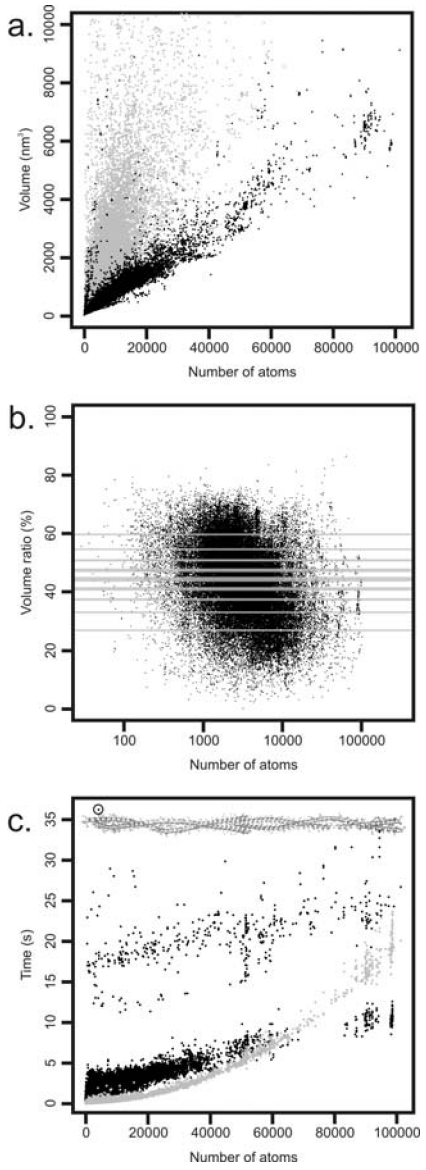

Figure 6. Squeeze-E efficacy. (a) Cell volume against protein size. Gray, rhombic dodecahedron; black, NDLP. (b) Volume ratios between NDLP and rhombic dodecahedra. The horizontal lines indicate the deciles, with the thick central line showing the median of the distribution. (c) Time required for determining a rhombic dodecahedron unit cell with Gromacs (gray) or an NDLP cell with Squeeze (black), using five subdivisions, against protein size. The inset shows the structure of a helix coil, for which the calculation of the NDLP takes an atypically long time (36 s, marked in the plot with a circle).

independent of other factors, and any other optimization can be combined with it.

From Figure 6b, it is also evident that for some structures the decrease amounts to more than 95%. These are mainly very elongated structures, in particular coiled-coils, such as the one shown in the inset in Figure 6c.

Figure 6c shows the times required for determination of the NDLP unit cell against protein size. Calculations were performed on one core of an Intel Core i7, 2.80 GHz. The maximum time taken for any structure is 36 s. This makes a comparison with the previous method to determine the NDLP futile, as these calculations typically took tens of minutes to hours, even prohibiting processing the entire PDB database. In fact, the time needed to calculate the NDLP is now more on par with the time required to calculate a rhombic dodecahedron, shown in gray, using the standard method of Gromacs. The NDLP timings show two regimes, both of which have a linear dependency on the system size. The difference between the regimes can be related to the shape and the symmetry of the protein. This is illustrated in the inset, which shows one of the most pathological cases, taking 36 s to calculate the NDLP for a protein with a modest number of atoms.

When calculating the packing for an ensemble of structures, the difference becomes much larger even (data not shown). Because the new implementation avoids the triangulation of point sets and the calculation of intersections, the computational cost is diminished.

As different approximations to the contact body are used by the two NDLP algorithms, it is worthwhile to compare the resulting minimal distance between any two periodic images to the distance criterion set. Both methods result in somewhat lower distances than specified (data not shown), which is caused by the triangulations being inscribed. The previous method results in minimal distances about 5% ($-5.15 \pm 2.01\%$) too small on a distance of 2.5 nm. In contrast, for the method proposed here, the difference amounts to about one-third of a percent ($-0.35 \pm 0.04\%$). The difference between the implementations can be explained from the use of a crude approximation to the sphere (five points) used for dilation of the molecule in the previous algorithm.

Taken together, the new algorithm is much faster, is more accurate, and handles ensembles properly and efficiently. In the following paragraphs, a number of examples are given for the use of the NDLP cell for ensembles, comparing the results to a rhombic dodecahedron unit cell.

The Ultimate Ensemble: 184 Structures of Phospholamban Solved by NMR. As a first test of the efficacy of our method, the largest ensemble in the PDB database (PDB ID: 2HYN[7]) was taken, which contains 184 configurations of the unphosphorylated human phospholamban pentamer. With over 800 000 atoms, this is one of the largest files available from the database.

Starting with a sphere approximation of 545 vertices (four subdivisions), the calculation of the NDLP with a minimal distance between images of 2.5 nm takes 12 s on an Intel Core i7 930 running at 2.80 GHz. The resulting simulation cell, shown in Figure 7a, has a volume of 543 nm^3, which would yield a total system size of 54 000 atoms when solvated. On the same machine, determining the NDLP with a sphere approximation of 2113 vertices (five subdivisions) takes 51 s and yields a slightly smaller simulation cell with a volume of 537 nm^3, which would result in a solvated system containing 53 500 atoms. With six subdivisions, the calculation time increases to 4 min, and the resulting volume is 522 nm^3.

A rhombic dodecahedron with the same distance criterion would yield a total volume of 650 nm^3, considering only a single model, and result in a solvated system of 64 700 atoms. Using the ensemble NDLP thus decreases the size of the system by 17−19%, while explicitly accounting for the flexibility that may be expected, based on the spread in the NMR models.

Ensemble Estimates from Molecular Simulations. Obviously, one of the most common ways to obtain an estimate of the conformational space accessible in a simulation is the simulation itself. When a simulation has been performed in a more conventional unit cell, the method proposed here

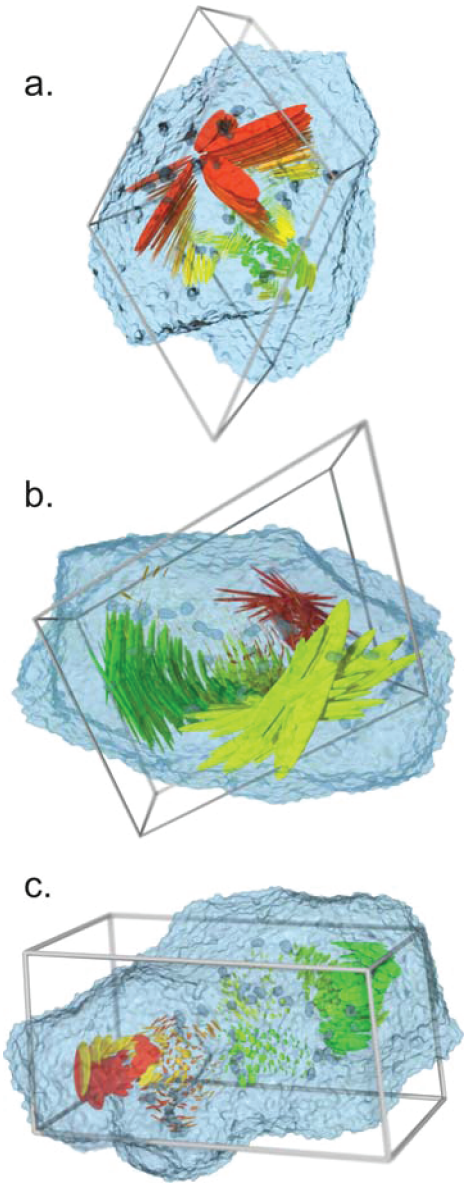

Figure 7. Optimal solutions for ensembles determined by Squeeze-E. (a) The optimal unit cell for the 184 member NMR ensemble of the human phospholamban pentamer (PDB ID: 2HYN[7]). The ensemble is represented as a set of spheroids, depicting the anisotropic C_α variances. The solvent is shown as the molecular shaped box representation, illustrating the distribution around the protein. (b) The optimal unit cell for the 100 member ensemble of the Lac Repressor (PDB ID: 1EFA[8]), as obtained from a 40 ns MD simulation. (c) The optimal unit cell for the 100 member ensemble of the hemagluttinin–hemagluttinin antibody complex (PDB ID: 2VIR[9]) as estimated by elastic network modeling. Again, the ensemble is represented as spheroids, depicting anisotropic C_α variances.

may be used to decrease the volume for replica simulations, while ensuring that all conformations sampled in the original simulation will fit in the new unit cell. In addition, this method can be used to resolvate a system in which interactions between periodic images occurred during the simulation. As an example, a simulation of the Lac Repressor was performed in its free form, unliganded and not bound to operator, initially using a NDLP cell in its original form.[1] It was found that during the simulation the DNA binding domains dispatched from the core part of the protein (unpublished data) and through the flexible linker regions made contact with a periodic image. Using the ensemble resulting from this simulation to set up a new simulation system using the method proposed here, it could be ensured that such large deformations would fit in the simulation cell for subsequent simulations, avoiding interactions over the periodic boundaries.

The simulation cell thus obtained, shown in Figure 7b, had a volume of 661 nm^3, yielding a total simulation size of 62 700 atoms. This volume is considerably larger than that obtained from the original structure, which measured 421 nm^3 and contained 38 600 atoms. However, a rhombic dodecahedron determined from the starting structure would already yield a volume of 1131 nm^3, corresponding to a total size of 109 300 atoms. In this case, the use of an ensemble NDLP unit cell thus still offers a reduction of 42%, thereby taking conformational changes observed previously into account.

Ensemble Estimates from Elastic Network Modeling. Although fairly robust, performing simulations to obtain ensemble estimates for setting up a more efficient system is rather expensive. An alternative approach is to estimate the conformational freedom from the elastic properties of the protein, using elastic network modeling (ENM) or related methods. Such runs typically take several hours and are easily performed using one of a number of ENM Web-based services. This approach is illustrated here by a simulation performed on an antibody–antigen complex.

Complexes of antibodies and antigens are typically large and nonregularly shaped. Whereas the size poses limitations on the simulations, the shape is expected to allow a significant decrease of the simulation volume when using an NDLP method. As an example, consider the complex of the influenza virus hemagluttinin in complex with a neutralizing antibody (PDB ID: 2VIR[9]). This complex contains a total of approximately 700 amino acids, and the radii of gyration about the *x*, *y*, and *z* axes are 3.3, 2.7, and 2.7 nm, exemplifying the elongated shape of the complex. Solvating this complex in a rhombic dodecahedron unit cell with a minimal distance between images of 2.0 nm would yield a system with a total size of approximately 230 000 atoms. To obtain an estimate for the conformational freedom of the complex, it was submitted to the elNémo server using default parameters.[10] From the results of that run, the extreme projections on the first 10 eigenvectors were used to form an ensemble, which was subsequently used as the input for calculating the ensemble packing. With five divisions for the sphere approximation and a distance between periodic images of 2.0 nm, this ensemble results in a system with a total volume of 540 nm^3, shown in Figure 7c, corresponding to a fully solvated system of approximately 49 000 atoms. This means that for this complex, the simulation volume can be decreased by almost 80%, while explicitly accounting for some conformational change.

■ DISCUSSION AND CONCLUSION

In the foregoing, a method has been described for determining an optimal unit cell for single structures and ensembles, for setting up molecular simulations with periodic boundary conditions.

This method started out as a routine optimization of an existing algorithm to handle ensembles more efficiently but resulted in new approaches to construct contact bodies and bodies of minimal distance, as well as a new method to determine the optimal packing from them, finally resulting in a complete revision of the original algorithm. The most notable

features are (1) a revised mathematical background, exemplifying the relation between the contact body and the body of minimal distance, as well as the possibility to construct those correctly for ensembles, (2) the use of a preset triangulation with fixed and well-defined topology that is scaled to the surface of minimal distance, and (3) the search for the NDLP from the discrete set of vectors forming the surface. These factors make the computational overhead of setting up an NDLP unit cell negligible.

The routines could possibly be optimized further for specific cases, such as symmetric protein assemblies, but that seems interesting mainly as an academic exercise, not yielding a benefit in a practical sense.

Taken together, the renewed NDLP method appears a suitable addition to the means for setting up and running simulations. A practical limitation is that the use of the NDLP unit cell requires an implementation of rotational constraints. The implementation of the constraints developed by Amadei et al.[3] has thus far been limited to Gromacs, in particular to version 3, and has only been used by a small number of groups. A port to Gromacs 4 has been hampered by the domain decomposition scheme,[11] in which molecules may be fragmented over processors and across box boundaries. Fortunately, a recent effort has enabled the reimplementation of the roto-translational constraints in the newest Gromacs version, without compromising efficiency (manuscript in preparation). This implementation has been made available through the MD Web server, developed within the WeNMR project, which also supports the use of the NDLP unit cell.

In addition to this practical issue, there are a few theoretical issues that require reflection. First of all, it has been shown previously that the unit cell is an integral part of the simulation conditions and may influence the resulting ensemble.[2] It could be possible that the information used to set up the NDLP unit cell in some way results in a "memory" of that information during the simulation, biasing the resulting ensemble. However, if that were true, then the same effect should have been seen in a more severe form in NDLP unit cells set up for a single structure. Yet the opposite appeared to be the case, with the NDLP unit cell yielding ensembles indistinguishable from those obtained in a rhombic dodecahedron, whereas a rectangular box and a truncated octahedron were seen to yield different ensembles. The explanation suggested was that the water distribution in the periodic system could play a role, which is independent of orientation in a rhombic dodecahedron and regularized in an NDLP unit cell. It is expected that the distance between periodic images does play a significant role in this respect, and future studies should resolve the dependence of results obtained in molecular simulations, in relation to the box type and the distance used.

A second concern with the use of an NDLP unit cell is that conformational changes may occur that are not covered by the input ensemble and thus still cause violation of the minimal distance criterion. This is particularly likely when using estimates from short simulations and extrapolations from ENM as input for long simulations of proteins with (partially) disordered regions or flexible linkers. Again, this is a concern that also applies to conventional setups but may be more severe when using an NDLP unit cell. As a general rule of thumb, one could argue that the longer the simulation is aimed to be, the more time one should spend thinking over the setup, including the choice and details of the periodic boundary conditions. If larger conformational changes or partial unfolding is to be expected, an option is to increase the distance between periodic images, which for an NDLP unit cell can usually be achieved while still yielding an overall reduction in system size.

For very flexible molecules, such as proteins with considerable unfolded/disordered regions, an additional issue may arise. The method described determines the contact body of the ensemble after least-squares fitting of each structure in the ensemble to a reference. For flexible molecules, the fitted ensemble is dependent on the reference, and thus ill-defined. However, if the rotational constraints of Amadei et al. are used, or a similar method which uses a reference structure to remove rotational degrees of freedom, then this does not pose problems, provided the same reference is used.

As a final note, the routines described here are aimed at optimizing molecular simulations, but they may also be of interest for other fields of science. In this regard, it is worth mentioning that contact bodies play a role in motion path planning for robots, and the routines presented could be developed further for such processes.

The method has been termed Squeeze-E for its efficacy in reducing the amount of solvent, with the E indicating its specific usability for ensembles. The method is available at http://haddock.chem.uu.nl/services/SQUEEZE/ and will be made available as a contribution to the Gromacs MD package (http://www.gromacs.org/). As mentioned, the use of this type of unit cell is also available as an option for the weNMR MD GRID server that is described elsewhere in this issue.

AUTHOR INFORMATION

Corresponding Author

*Phone: +31.(0)50-3634336. Fax: +31.(0)50-3634800. E-mail: tsjerkw@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the European Union FP6 STREP "BacAbs MS" (Grant Number: LSHB-CT-2006-037325) and the European Union FP7 I3 project "e-NMR" (contract number 213010-e-NMR).

DEDICATION

This work is dedicated to Wilfred van Gunsteren, in honor of his 65th birthday. He is among the people that have shaped molecular dynamics to become the field it nowadays is. Not only has he made many contributions, he has also always been keen to see advances and stimulate developments from others. This was particularly true during the informal BIOMOS meetings, held in Burg Arras, where new insights and new viewing angles were invariably met with enthusiasm, with invitation to continue and show more. We thank Wilfred for the discussions and collaboration.

REFERENCES

(1) Bekker, H.; van den Berg, J. P.; Wassenaar, T. A. *J. Comput. Chem.* **2004**, *25*, 1037−1046.

(2) Wassenaar, T. A.; Mark, A. E. *J. Comput. Chem.* **2006**, *27*, 316−325.

(3) Amadei, A.; Chillemi, G.; Ceruso, M. A.; Grottesi, A.; Di Nola, A. *J. Chem. Phys.* **2000**, *112*, 9−23.

(4) Edelsbrunner, H.; Mucke, E. P. *ACM Trans. Graph.* **1994**, *13*, 43−72.

(5) Clemons, W. M., Jr.; Davies, C.; White, S. W.; Ramakrishnan, V. *Structure* **1998**, *6*, 429−438.
(6) Boal, N.; Domínguez, V.; Sayas, F. J. *J. Comput. Appl. Math.* **2008**, *211*, 11−22.
(7) Potluri, S.; Yan, A. K.; Chou, J. J.; Donald, B. R.; Bailey-Kellogg, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 203−219.
(8) Bell, C. E.; Lewis, M. *Nat. Struct. Biol.* **2000**, *7*, 209−214.
(9) Fleury, D.; Wharton, S. A.; Skehel, J. J.; Knossow, M.; Bizebard, T. *Nat. Struct. Biol.* **1998**, *5*, 119−123.
(10) Suhre, K.; Sanejouand, Y.-H. *Nucleic Acids Res.* **2004**, *32*, W610−614.
(11) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.