
A New Program for Optimizing Periodic Boundary Models of Solvated Biomolecules (PBCAID)

XIAOLIANG QIAN, DANIEL STRAHS, TAMAR SCHLICK

Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and Howard Hughes Medical Institute, 251 Mercer Street, New York, New York 10012

Received 13 December 2000; accepted 22 May 2001

ABSTRACT: Simulations of solvated macromolecules often use periodic lattices to account for long-range electrostatics and to approximate the surface effects of bulk solvent. The large percentage of solvent molecules in such models (compared to macromolecular atoms) makes these procedures computationally expensive. The cost can be reduced by using periodic cells containing an optimized number of solvent molecules (subject to a minimal distance between the solute and the periodic images). We introduce an easy-to-use program "PBCAID" to initialize and optimize a periodic lattice specified as one of several known space-filling polyhedra. PBCAID reduces the volume of the periodic cell by finding the solute rotation that yields the smallest periodic cell dimensions. The algorithm examines rotations by using only a subset of surface atoms to measure solute/image distances, and by optimizing the distance between the solute and the periodic cell surface. Once the cell dimension is optimized, PBCAID incorporates a procedure for solvating the domain with water by filling the cell with a water lattice derived from an ice structure scaled to the bulk density of water. Results show that PBCAID can optimize system volumes by 20 to 70% and lead to computational savings in the nonbonded computations from reduced solvent sizes. © 2001 John Wiley & Sons, Inc. *J Comput Chem* 22: 1843–1850, 2001

Keywords: periodic boundary conditions; particle-mesh Ewald; molecular dynamics; space-filling polyhedra; solvation

Correspondence to: T. Schlick; e-mail: schlick@nyu.edu

Contract/grant sponsor: NIH; Contract/grant number:

GM 55164

Contract/grant sponsor: NSF; Contract/grant numbers: BIR-

94-23827EQ and ASC-9704681

Contract/grant sponsor: John Simon Guggenheim Fellowship (to T.S.)

Introduction

State-of-the-art molecular dynamics (MD) and Monte Carlo simulations for biomolecules represent solvent molecules explicitly for accurate modeling of equilibrium properties and sampling of conformations. Solute macromolecules, although principally affected by nearby solvent molecules, require a sufficiently large solvent layer to eliminate artifacts associated with the absence of bulk solvent. In periodic boundary conditions (PBC) and particle-mesh Ewald (PME) methods for long-range interactions, in particular,^{1,2} additional solvent molecules are required to fill the volume of the periodic cell; the periodic cell is then translated and/or rotated symmetrically along the defining lattice vectors to fill one (PBC) or more (PME) layers of space-filling polyhedra. Both PBC and PME methods can suffer from the artificial periodicity when small periodic domains are used.³ For example, an alanine polypeptide (18 Å size) simulated with PME was recently reported to be superficially stabilized in an α -helical geometry by a regular periodic cube with a 30 Å edge.³ Solvated cells with solute-image distances larger than those discussed above can also suffer from artifacts if the solute rotates to align one of its longer axes along a shorter periodic cell axis.

The computational cost of simulating such large periodic cells is considerable because a large percentage of the system size reflects solvent atoms; thus, it is important to reduce the cell size as much as possible while maintaining a reliable model. Mezei has shown that optimal solute rotations can increase the minimal distance between periodic images, thus permitting the periodic cell size to be reduced.⁴ The associated decrease in the cell volume and solvent number has been estimated to be ≈ 10 –30%, leading to even greater overall computational savings (because cost scales approximately as $n^{1.5}$, where n is the number of atoms). The procedures described by Mezei (as well as protocols to optimize solute placement in a spherical droplet⁵) are included in the molecular simulation support package SIMULAID (available from <http://fulcrum.physbio.mssm.edu/~mezei/>).

Here we introduce new variations to increase the efficiency of the periodic cell optimization methods pioneered by Mezei and introduce new capabilities to biomolecular simulation. In contrast to SIMULAID, which explicitly calculates all solute/image distances, our program PBCAID adopts the efficient procedure of limiting the calculations to the solute's surface atoms and computing distances between

these surface atoms and the faces of the periodic cell. We also limit rotations to the periodic cell vertices (i.e., not atoms), and can model seven known space-filling periodic cells, as shown in Figure 1. In addition to optimizing the periodic cell dimension, PBCAID includes a protocol for solvating the cell with water using an ice lattice scaled to reproduce the bulk density of liquid water. PBCAID's optimization algorithm is simpler and, hence, more efficient for a given periodic shape, and allows users to choose the most optimal or suitable domain for the application at hand. Future versions of PBCAID might expand the target optimization function and address the issue of reducing simulation artifacts due to the use of periodic cells.

Method

PROGRAM OVERVIEW

The choice of periodic domain shape depends on two factors. The domain must suit the long-range energy approximation used (e.g., PME methods prefer simple integer lattices, such as rectangular prisms and rhombic dodecahedron). At the same time, the domain and embedded model should lead to an overall efficient protocol (i.e., the smallest system size) and not introduce artifacts (i.e., alignment of dipolar moments). As previously shown,⁴ the solute orientation inside a periodic cell can be optimized to minimize the periodic cell volume while simultaneously maximizing the minimum solute/image distance. The minimum solute/image distance D_{ij} for a given orientation is defined as the minimum of all atom pair distances between the solute and all images generated by rotating and/or translating the primary cell through its periodic transformations (see Fig. 2, top). During the course of SIMULAID's optimization of the solute rotation, the user has an option to discard internal atoms (defined as three- and four-bonded atoms). The option to discard internal atoms results in a 2D-limiting surface to the 3D-macromolecule; it can be empirically shown that $n^{2/3}$ atoms are retained where n is the number of atoms. The distance calculations of SIMULAID are thus of order $\mathcal{O}(n^{4/3})$.

In practice, the minimum solute/image distance is obtained for atom pairs at or near the surface of the solute. PBCAID thus enhances efficiency by defining the surface atoms of the solute using a spherical grid. The spherical grid is generated with latitude/longitude lines spaced at regular intervals, such as $m = 5^\circ$. Within each solid angle formed

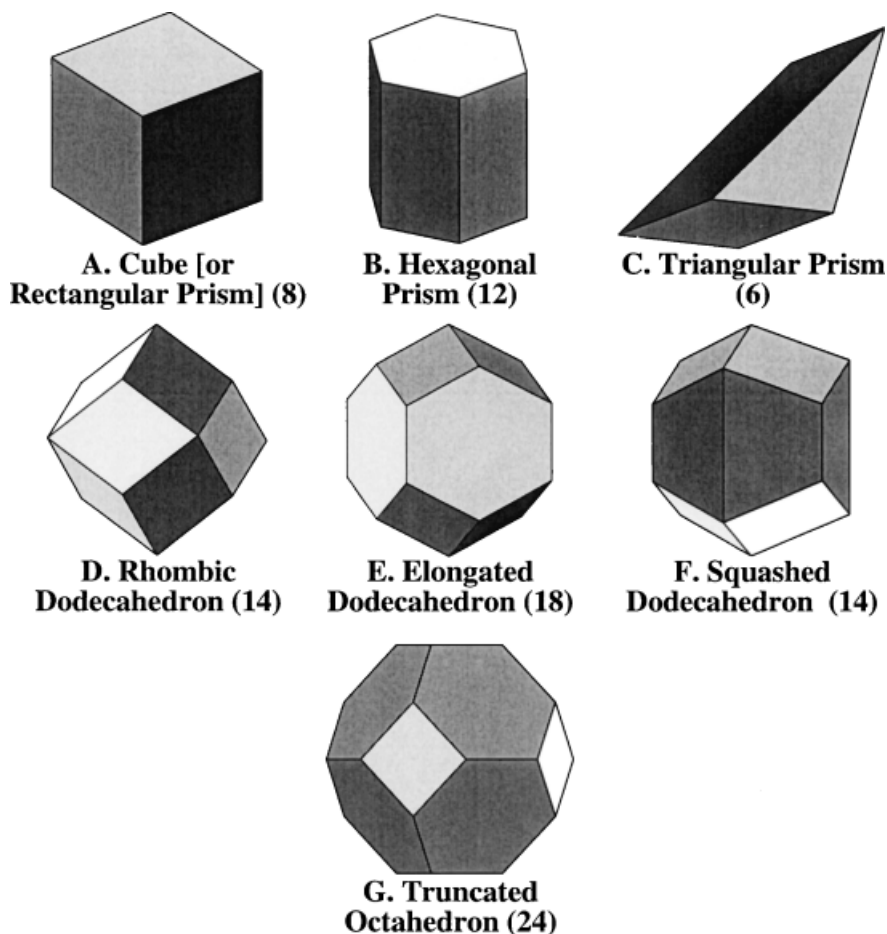


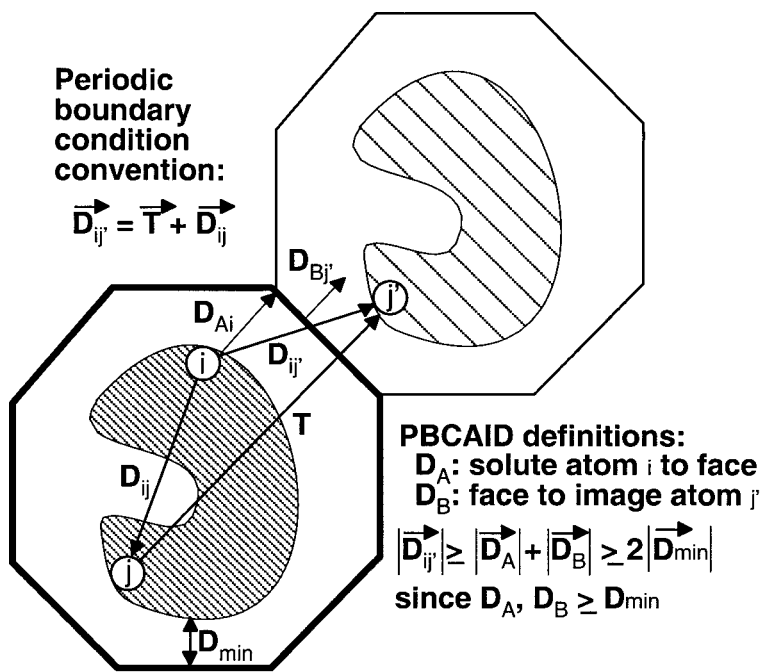
FIGURE 1. Space-filling polyhedra included in PBCAID. The number of vertices for each domain is indicated.

by this grid, only one atom with the largest radial distance from the origin is identified as a surface atom for our calculations. The total number of surface atoms is equal to $(360/m) \cdot (180/m)$. Thus, for a 5° grid interval, no more than 2592 atoms are used to define the surface. This discretization of the solute surface reduces the computational size significantly without sacrificing accuracy.

The efficiency of the optimization procedure can be further improved by computing only the minimal distance to the face of the periodic cell, rather than the distance to the image atoms. The distances D_{ij} between any solute atom i and any image atom j' are of the form: $\vec{D}_{ij'} = \vec{D}_{ij} + \vec{T}$, where \vec{D}_{ij} is the vector from atom i to j in the solute, and \vec{T} is a transform vector defining the image cell (Fig. 2, top). The distance $\vec{D}_{ij'}$ satisfies the triangle inequality: $\vec{D}_{ij'} \geq \vec{D}_{Ai} + \vec{D}_{Bj'} \geq 2\vec{D}_{\min}$, where \vec{D}_{Ai} is the minimum distance from atom i to the face of the cell closest to atom j' , and $\vec{D}_{Bj'}$ is the corresponding distance for the face closest to atom j' within

the transformed image. We define \vec{D}_{\min} as the minimum distance from any surface atom of the solute to any face of the periodic cell; thus, $\vec{D}_{\min} \leq \vec{D}_{Ai}$ for any distance measured between an atom i and any face of the periodic cell. Computing D_{\min} is an $\mathcal{O}(\alpha n)$ procedure, where α is the number of faces of the periodic cell. Because α is a constant typically 100 times smaller than the number of atoms, this procedure is more efficient than the $\mathcal{O}(n^{4/3})$ procedure used by SIMULAID. In sum, our calculation optimizing the solute/image distances involves computing distances using only surface atoms and computing the minimal distance to the face of the periodic cell, rather than the image atoms.

A user-specified distance $D_{\text{target}} \leq D_{\min}$ defines the minimal distance between the solute and the faces of the periodic cell that we tolerate (e.g., 10 \AA water layer); the initial cell size is also determined by the sum of D_{target} and the solute projection along the three orthogonal axes. PBCAID rotates only the periodic cell vertices, indicating that no more than 24 coordinates need to be rotated, further reducing



Distance from an atom i to a periodic face:

$$D = \frac{(\vec{a} / |\vec{a}|) \times (\vec{b} / |\vec{b}|)}{\cdot (\vec{r}_i - \vec{v}_l)}$$

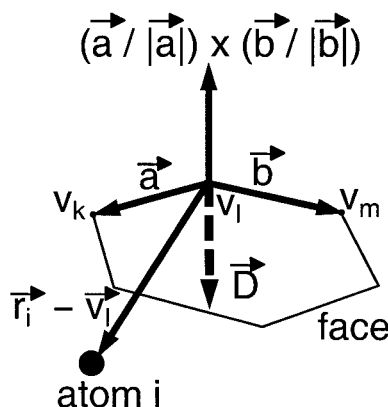


FIGURE 2. Illustration of the vectors and distances used in developing the PBCAID algorithm. Top: An irregularly shaped solute (shaded area) is indicated in an octagonal cell, approximating a 2D projection of the truncated octahedron. Atoms i and j , and the corresponding image atom j' (at upper right) are shown, with vectors indicating the transformations and distances between selected points. Bottom: Vertices (v_k, v_l , and v_m) and vectors (\vec{a}, \vec{b} , and $(\vec{r}_i - \vec{v}_l)$) as defined in the text are used to calculate the distance $D(i, f)$ from atom i to a face f of the periodic cell. The relationship between these vectors and vertices for a hexagonal face (resembling the hexagonal prism or elongated dodecahedron) is shown.

the computational cost. After the optimal cell rotation is obtained by minimizing $D_{ij'}$, the inverse of the optimal rotation matrix specifies the optimal rotation to apply to the solute, with respect to the original cell orientation.

OPTIMIZATION DETAILS

Let v_k be the coordinates of a vertex of the cell; for a given periodic domain, there are K ver-

tices. K ranges between 6 (for the triangular prism) to a maximum of 24 (truncated octahedron); see Figure 1. The orientation of the periodic cell is defined by the quaternion rotation matrix $R(e_0, e_1, e_2, e_3)$:

$$R(e_0, e_1, e_2, e_3) = \begin{pmatrix} e_0^2 + e_1^2 - e_2^2 - e_3^2 & 2(e_1e_2 + e_0e_3) & 2(e_1e_3 - e_0e_2) \\ 2(e_1e_2 - e_0e_3) & e_0^2 - e_1^2 + e_2^2 - e_3^2 & 2(e_2e_3 - e_0e_1) \\ 2(e_1e_3 + e_0e_2) & 2(e_2e_3 - e_0e_1) & e_0^2 - e_1^2 - e_2^2 + e_3^2 \end{pmatrix}$$

where e_0, e_1, e_2 , and e_3 satisfy the relation that $\sum_i e_i^2 = 1$ and are known as Euler (or Cayley–Klein) rotation parameters.⁶ The set of rotated cell vertices \vec{V} is related to the initial set \vec{V}_0 by:

$$\vec{V} = \mathbf{R}(e_0, e_1, e_2, e_3)\vec{V}_0. \quad (1)$$

The distance between any solute surface atom i ($i \in S$), and a face f of the periodic cell ($f \subset F$), can be defined by three vertices of the face: v_f . We define vectors \vec{a} and \vec{b} as: $\vec{a} = \vec{v}_k - \vec{v}_l$, $\vec{b} = \vec{v}_m - \vec{v}_l$, where \vec{v}_k , \vec{v}_l , and \vec{v}_m are vectors from the origin to the three vertices of $\{v_f\}$. The cross product between $\vec{a}/|\vec{a}|$ and $\vec{b}/|\vec{b}|$ thus defines the unit normal to the face (Fig. 2, bottom). The distance $D(i, f)$ between atom i and the cell face f is the projection of the unit normal vector against the vector from any vertex of the face f to atom i :

$$D(i, f) = \left(\frac{\vec{a}}{|\vec{a}|} \times \frac{\vec{b}}{|\vec{b}|} \right) \cdot (\vec{r}_i - \vec{v}_l), \quad (2)$$

where \vec{r}_i defines the coordinates of solute atom i . The minimal distance D_{\min} is the minimum distance between all atoms i and all faces f for a given orientation $\mathbf{R}(e_0, e_1, e_2, e_3)$: $D_{\min} = \min_{i \in S, f \subset F} D(i, f)$. For any particular rotation $\mathbf{R}(e_0, e_1, e_2, e_3)$, though there is only one value of D_{\min} , more than one surface atom may have values of $D(i, f) = D_{\min}$.

We optimize the scoring function $F(e_0, e_1, e_2, e_3, a, b, c)$:

$$F = \sigma_1 \left(\frac{D_{\min} - D_{\text{target}}}{D_{\text{target}}} - \delta \right)^2 + \sigma_2 \left(\sum_{i=0}^3 e_i^2 - 1 \right)^2 + \sigma_3 (V(a, b, c)) \quad (3)$$

where δ is a tolerance (≈ 0.01); $\{e_i\}$ for $i = 0, 1, 2, 3$ are the quaternion rotation parameters; $V(a, b, c)$ is the volume of the periodic cell for the dimensions a, b , and c ; and σ_1, σ_2 , and σ_3 are adjustable parameters (we use $10^8, 10^8$, and 1 , respectively). Note that D_{\min} , a function of all the distances $D(i, f)$, is also implicitly a function of $R(e_0, e_1, e_2, e_3)$; thus, the distances $D(i, f)$ are implicitly considered in the scoring function. The three terms of the objective function F are designed to successively reduce the cell dimensions towards target distance D_{target} , restrain the quaternion parameters to represent a rotation matrix, and minimize the overall volume of the periodic cell.

Other formulations of the objective scoring function can be used; for example, the position of the cell center might be optimized or the adjustable parameter σ_3 might be scaled to reflect an intrinsic dependence on the system size (M. Mezei, personal communication). The quaternion rotation parameters can also be constrained through a separate

Lagrange multiplier method (M. Mezei, personal communication), suggesting that random seeds can be used to initialize different optimization cycles and efficiently search the periodic cell rotations.

We use the simplex method for this discrete optimization problem.⁷ For an n dimensional optimization problem, the simplex method requires $n + 1$ initial conditions. These initial conditions include the initial orientation and dimensions of the cell, which are defined by the four quaternion rotation parameters e_i and the three lengths a, b , and c . The rotations specifying the initial quaternion parameters are defined randomly.

After the optimization is complete, we apply \mathbf{R}^{-1} , the inverse of the current rotation matrix \mathbf{R} , to the vertices \vec{V} and the solute atoms \vec{X} to define the optimal solute orientation: $(\mathbf{R}^{-1}\vec{V}, \mathbf{R}^{-1}\vec{X}) = (\vec{V}_0, \mathbf{R}^{-1}\vec{X})$, which follows from eq. (1).

Initially, we place the solute in the center of the cell. Although the solute orientation is not varied during the optimization procedures, the program has the option to initialize the solute orientation using either the original solute orientation or an orientation defined such that the largest gyration direction (or the principal axis) is oriented along the x -axis. The initial cell dimensions are set to be the minimal and maximal dimensions of the molecule along each coordinate axis plus D_{target} , the requested minimal distance to the cell boundary. The cell faces are defined by the chosen periodic domain, from among the seven illustrated in Figure 1.

SOLVATING PERIODIC CELLS

The solvation procedure uses the Ih phase hexagonal ice structure.⁸ An ice lattice is generated by symmetry operations applied to a four-oxygen subunit of the hexagonal ice cell; both translations and rotations are applied to this four-oxygen subunit to regenerate the hexagonal ice lattice. Because water (1 g/cm^3) is denser than ice (0.93 g/cm^3), the oxygen/oxygen atomic distance of the hexagonal ice structure (2.75 \AA) is scaled to be 2.6881 \AA , to reproduce the bulk density of water. Hydrogen atoms are added along the vectors connecting adjacent oxygen atoms in the hexagonal ice cell. Additional procedures are required to reproduce the bulk density of water: small random rotations and translations (not exceeding 12° and 2 \AA , respectively) are applied to each lattice cell, perturbing the lattice periodicity. Because the rotations and translations can place oxygen molecules outside the cell, an additional layer of ice cells is generated outside the periodic

TABLE I.
Optimization of Three Different Macromolecular Systems for: 14-bp DNA, TATA Element DNA Bound to the TATA-Box Binding Protein (TBP), and a DNA Primer/Template Segment Bound to Polymerase β (pol β).

Solute	Periodic Cell	Initial Volume/ Water Number	Final Volume/ Water Number	Percent Decrease	Min. Distance
[14 bp DNA]	sqdod (F)	388763 / 13007	110992 / 3731	71	20.6
	hex (B)	172784 / 5775	112892 / 3778	35	22.3
	cube (A)	156505 / 5226	129727 / 4299	17	21.8
	troct (G)	306559 / 10245	177628 / 5970	42	22.4
	rhomb(D)	281593 / 9406	204592 / 6847	27	25.3
[DNA/TBP]	elongd (E)	579458 / 19371	315428 / 10528	46	21.1
	rhomb (D)	434830 / 14517	326268 / 10908	25	20.9
	troct (G)	473383 / 15869	338716 / 11348	28	22.1
	sqdod (F)	1077302 / 36102	339624 / 11357	68	21.0
	hex (B)	478801 / 16001	386555 / 12836	19	25.4
	cube (A)	507102 / 16965	406410 / 13626	20	21.9
[DNA/pol β]	elongd (E)	880248 / 29413	403459 / 13490	54	20.3
	rhomb (D)	684528 / 22874	403855 / 13495	41	20.4
	cube (A)	683476 / 22868	412971 / 13805	40	21.4
	troct (E)	745220 / 24900	434132 / 14507	42	22.2
	sqdod (F)	1434416 / 47935	447299 / 14979	69	19.7

For each system, the periodic cells are ordered by increasing final volume (given in \AA^3). The minimum solute/image distance ($\min D_{ij}$) indicated in the last column is given in \AA for the optimized system. The space-filling polyhedra are labeled as in Figure 1.

box to permit the random rotations and translations to return oxygen atoms to the periodic box.

After the full lattice is generated, oxygen atoms are tested to exclude those placed outside the primary periodic cell. The program incorporates a procedure to exclude waters that violate a user-specified distance from the solute atoms; however, packages like CHARMM⁹ or AMBER¹⁰ possess sophisticated exclusion algorithms, and users may choose to resort to those routines instead.

Results and Discussion

The algorithms described are collected in our software package, PBCAID (<http://monod.biomath.nyu.edu/index/software/PBCAID/index.html>). Table I details results for three systems: a 14-bp DNA containing an A-tract variant of the adenovirus major late promoter TATA-box (888 solute atoms),¹¹ human TATA-box binding protein (TBP) complexed with the AdMLP TATA element (3952 solute atoms),¹² and human DNA polymerase β (pol β) complexed with primer and template (6418 solute atoms).¹³ For each system, we describe the results of optimizing several space-filling shapes. Examples of the initial and

final states of the two most efficient periodic shapes are shown in Figure 3 for each system, indicating the solute orientation and the water molecules placed by our solvation procedure.

In each case, the solvent layer distance D_{target} is set to 10 \AA . The solutes are initially oriented by aligning the principal geometric axis along the x -axis and the next-largest axis along the y -axis. We see that the percent decrease in the number of water molecules is greater than 10%, indicating savings on the order of more than 1000 water molecules. If we assume that the computational cost scales as $n^{1.5}$, where n is the number of atoms, the optimized cells can lead to an increase of efficiency of 40% or more relative to the initial solute orientation (assuming a final volume decrease of 30%).

Traditional periodic cell shapes, such as the hexagonal prism for DNA, have often been used in simulations; such shapes yield smaller final volumes, and may be preferred if artifacts (such as alignment of dipolar moments) are not significant. Interestingly, the cube (rectangular prism) often used in simulations of DNA with PME can be optimized to be only slightly worse than the hexagonal prism, yielding a final volume 16% larger.

The final choice of a candidate periodic domain relies on several factors, such as the method used

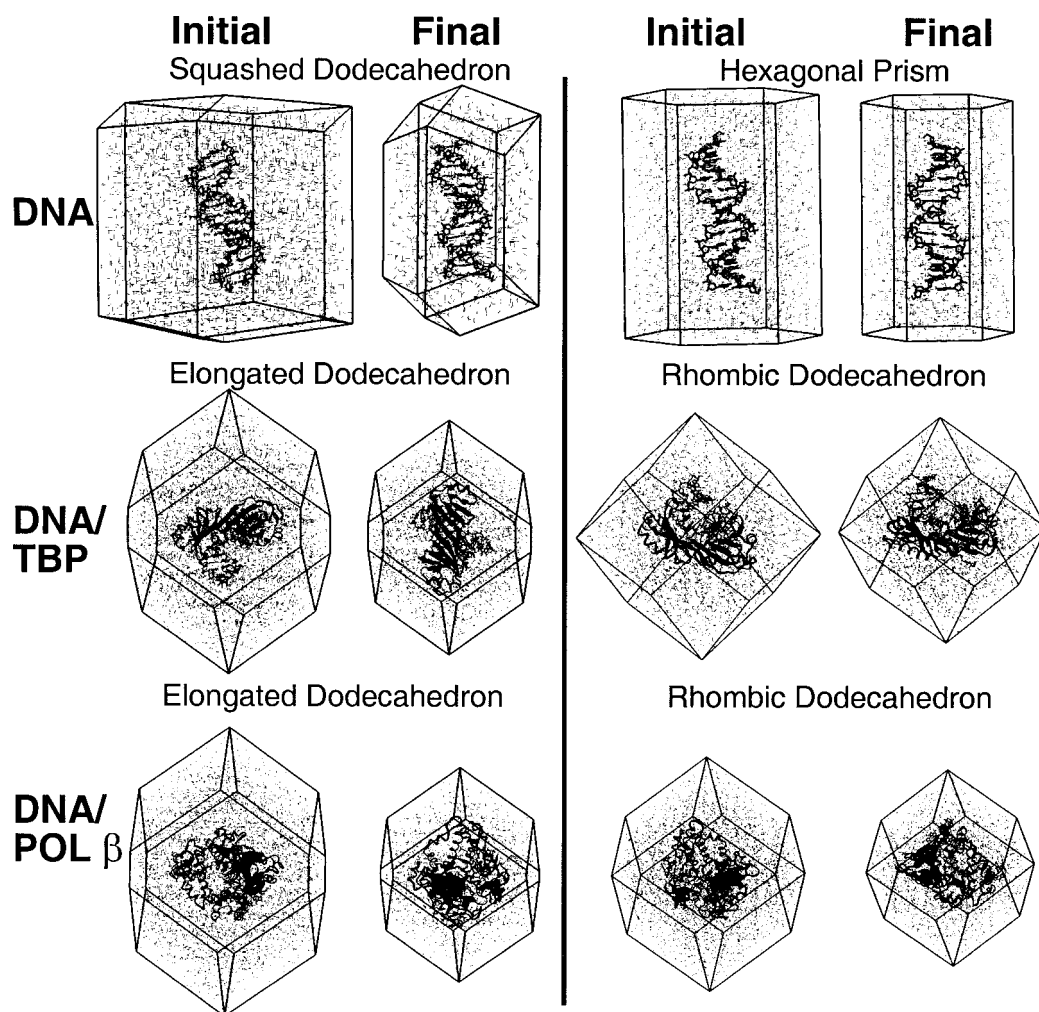


FIGURE 3. Examples of optimized polyhedra of three systems: 14-bp DNA, DNA/protein complex of the TATA element bound to the TATA-box binding protein (TBP), and DNA/protein complex of a primer/template DNA bound to polymerase β ($\text{pol } \beta$). For each system, examples are shown for the initial and final size of the two most efficient polyhedra. In each case, we show the solute molecule in the initial and optimized orientation, and the water molecules placed within the cell by our solvation procedure. See Table I for cell and percent decrease following optimization.

for calculating the long-range nonbonded interactions, the desired thermodynamic ensemble, and whether or not center-of-mass rotations are suppressed. Still, the computational costs associated with choosing nonoptimal cells for particular methods may be mitigated by our optimization methods. For example, the squashed dodecahedron, like the rhombic dodecahedron, is derived from close-packing of spheres,¹⁴ suggesting that this domain has a comparable best volume/inscribed sphere ratio like a rhombic dodecahedron, and should be very suitable to globular proteins. Although most globular proteins are clearly not ideally spherical, quasi-spherical periodic cells are still preferred if all possible orientations of the solute macromole-

cule are sampled. Practical considerations of non-idealities and the limited sampling available from most simulation techniques suggests that the use of quasi-spherical periodic cells must be balanced with other periodic cell shapes.

Alternatively, the use of some of the periodic cells we describe here may have special benefits, because some of the periodic systems require translational and *rotational* transforms (such as the squashed dodecahedron). Interestingly, the two ends of the squashed dodecahedron composed of parallelograms are identical; this identity indicates that the images bounding these two ends have the same slip plane translations and rotations, and thus do not have inverse images. Inverse images indicate

that for each image, there is another image whose location is described by the opposite translation/rotation. Because the positions of the inverse image atoms are identical to the first image, several modeling programs (such as CHARMM⁹) use only one of a pair of inverse-related images to calculate the periodic domain, resulting in a computational saving. While the lack of these inverse images leads to increases in computational time, the periodicities associated with the squashed dodecahedron are quite different from the artificial periodicities of standard domains, and may have interesting properties.

Finally, we note that Bekker has described an interesting alternative to optimizing solute orientation within periodic polyhedra.¹⁵ In Bekker's approach, all space-filling polyhedra are equivalent, because they fill three-dimensional (3D) space. Solute orientation can be optimized to determine the most efficient packing by placing four copies of the solute's center of mass at the vertices of a tetrahedron, and then optimizing the volume of the tetrahedron. Bekker suggests using this optimized tetrahedra to tessellate 3D space with cubes; we note that other space-filling polyhedra work equally well and can be used.

PBCAID, written in the C++ programming language, uses the OpenGL graphics library to allow interactive control of the periodic cell optimization process. Future versions of PBCAID may include routines for different alignments of periodic cells, and alternative functions to treat optimally molecular systems with macroscopic dipoles.

Acknowledgments

We thank Prof. Mihaly Mezei for generously providing the SIMULAID program, helpful discussions, and inspiring this research. T. Schlick is an investigator of the Howard Hughes Medical Institute.

References

1. York, D. M.; Yang, W.; Lee, H.; Darden, T.; Pedersen, L. G. *J Am Chem Soc* 1995, 117, 5001.
2. Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J Am Chem Soc* 1995, 117, 4193.
3. Weber, W.; Hunenberger, P. H.; McCammon, J. A. *J Phys Chem B* 2000, 104, 3668.
4. Mezei, M. *J Comput Chem* 1997, 18, 812.
5. Mezei, M. *Inf Q Comp Sim* 1993, 37.
6. Goldstein, H. *Classical Mechanics*; Addison-Wesley Publishing Co.: Reading, MA, 1981; p. 148, 2nd ed.
7. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: Cambridge, MA, 1992, p. 408; <http://www.nr.com/>.
8. Hobbs, P. V. *Ice Physics*; Clarendon Press: Oxford, 1974.
9. Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J Comput Chem* 1983, 4, 187.
10. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G. L.; Kollman, P. A. *Comp Phys Commun* 1995, 91, 1.
11. Qian, X.; Strahs, D.; Schlick, T. *J Mol Biol* 2001, 308, 681.
12. Qian, X.; Strahs, D.; Schlick, T. work in progress, 2001.
13. Yang, L.; Wilson, S. H.; Broyde, S.; Schlick, T. submitted, 2001.
14. Steinhaus, H. *Mathematical Snapshots*; Oxford University Press: New York, 1983.
15. Bekker, H. *J Comput Chem* 1997, 18, 1930.