Journal of Mathematical Chemistry, Vol. 42, No. 4, November 2007 (© 2006)
DOI: 10.1007/s10910-006-9079-8

A central partition of molecular conformational space. III. Combinatorial determination of the volume spanned by a molecular system

Jacques Gabarro-Arpa

Ecole Normale Supérieure de Cachan, LBPA,CNRS UMR 8113, 61, Avenue du Président Wilson, 94235 Cachan Cedex, France
E-mail: jga@infobiogen.fr

Received 5 December 2005; revised 8 February 2006

In the first work of this series (Gabarro-Arpa, Comp. Biol. Chem. 27 (2003) 153–159) it was shown that the conformational space of a molecule could be described to a fair degree of accuracy by means of a central hyperplane arrangement. The hyperplanes divide the space into a hierarchical set of cells that can be encoded by the face lattice poset of the arrangement. The model however, lacked explicit rotational symmetry, which made impossible to distinguish rotated structures in conformational space. This problem was solved in a second work (Gabarro-Arpa, *Proc. 26th Ann. Int. Conf. of the IEEE EMBS* (San Franciso, 2004) 3007–3010) by sorting the elementary 3-dimensional components of the molecular system into a set of morphological classes that can be properly oriented in a standard 3-D reference frame. This also made possible to find a solution to the problem that is being addressed in the present work: *for a molecular system immersed in a heat bath we want to enumerate the subset of cells in conformational space that are visited by the molecule in its thermal wandering*. If each visited cell is a vertex on a graph with edges to the adjacent cells, here it is explained how such graph can be built.

KEY WORDS: molecular conformational space, hyperplane arrangement, face lattice, molecular dynamics

Mathematics subject classification: 52B11, 52B40, 65Z05

PACS: 02.70.Ns

1. Introduction

Molecular dynamics simulations (MDS) are an essential tool for the modeling of large and very large molecules, it gives us a precise and detailed view of a molecule's behaviour [1]. However, it has two limitations that hamper many practical applications: it is a random algorithm, as such it does not perform a systematic exploration of molecular conformational space (CS); and that currently,

the output from an MDS represents only a very small fraction of the volume spanned by the system in CS.

Here it is presented a complementary approach that locally is less precise but that can encompass a broader view of CS. It consists in dividing the CS into a finite set of cells, so that the only knowledge we seek about the system is whether it can be located in a given cell or not.

As was extensively discussed in [2] the partition is a variant of the $\mathcal{A}_N$ partition [3, 4]: a central[1] arrangement of hyperplanes that divides CS into a set of cells shaped as polyhedral cones, such that for a molecule with N atoms we have $(N!)^3$ cells. The set of hyperplanes is also a Coxeter reflection arrangement: the arrangement is invariant upon reflection on any of the hyperplanes.

This structure has three important properties as follows [2]:

1. Associated with a Coxeter arrangement there is a polytope [4] whose symmetry group is the reflection group of the arrangement. The face lattice poset[2] of the polytope is a hierarchichal combinatorial structure that enables us to manage the sheer complexity of CS, since with simple codes we can describe from huge regions down to single cells.

2. The information needed to encode any face in the polytope is a sequence of $3 \times N$ integers, which is a generalization of a structure known to combinatorialists as noncrossing partition sequence [4, 5].

3. The construction is modular: if we consider the CS of two disjoint subsets of atoms from a system, the CS of the union set has an associated polytope that is the cartesian product of the polytopes[3] of the two subspaces, and its partition sequence is the ordered union of the two partition sequences [2].

The last one is particularly important since the CS of the whole system can be built from that of the parts, and the CS of a small number of atoms is very much smaller than that of the whole molecule and we can reasonably assume that it can be thoroughly explored by an MDS. Moreover, in merging the CSs corresponding to subsets of atoms the number of cells grows exponentially while the length of coding sequences grows only linearly.

Partitions of higher dimensional spaces are widely used in physics (see [6] and references therein), hierarchical partitioning of conformational space has been used in chemistry to build topological representations of potential energy surfaces [7] (and references therein). This work is about the construction of a modular hierarchical partitioning of molecular CS.

---

[1] That pass through the origin.

[2] The faces in the induced decomposition of the polytope ordered by inclusion.

[3] If $P \subset \mathbb{R}^p$ and $Q \subset \mathbb{R}^q$ are polytopes the product polytope $P \times Q$ has the set of vertices $(x, y) \in \mathbb{R}^{p+q}$, where x and y are vertices of P and Q, respectively.

2. The basic construction

Let $(e_1, \ldots, e_N)$ be the standard basis in $\mathbb{R}^N$, the convex hull of the endpoints of the vectors $\{e_\mathbf{i}\}$ is a regular $(N-1)$-simplex : this gives a segment, an equilateral triangle and a tetrahedron in 2, 3, and 4-dimensions, respectively.

For each edge of the regular $(N-1)$-simplex there is an hyperplane $H_{\mathbf{ij}}$: $x_\mathbf{i} - x_\mathbf{j} = 0$, perpendicular to the edge and containing the other vertices, this hyperplane divides $\mathbb{R}^N$ in three regions. A point x can be in one of these as follows:

- $x_\mathbf{i} > x_\mathbf{j}$ the positive side, where the **i**th coordinate dominates the **j**th coordinate,
- $x_\mathbf{i} < x_\mathbf{j}$ the negative side, whith the **j**th the coordinate dominating the **i**th coordinate,
- $x_\mathbf{i} = x_\mathbf{j}$ on the plane.

This leads to a sign vector S for every point $x \in \mathbb{R}^N$, where the αth component $X_\alpha \in \{+, -, 0\}$ denotes wether x is on the positive side of H_α, on its negative side or lies on H_α.

Also notice that the line $x_1 = x_2 = \cdots = x_{N-1} = x_N$ is contained in every plane $H_{\mathbf{ij}}$, the orthogonal complement to this line is $\mathcal{U}: x_1 + x_2 + \cdots + x_{N-1} + x_N = 0$, on it we can define a partition, known to combinatorialists as $\mathcal{A}_{N-1}$ [3, 4], with the set of hyperplanes $\mathcal{H}_{\mathbf{ij}} = \mathcal{U} \cap H_{\mathbf{ij}}$. For reasons that are explained below the points outside $\mathcal{U}$ are not relevant to our construction.

The set of all points $x \in \mathcal{U}$ having the same sign vector S form a cell in the decomposition of $\mathcal{U}$ induced by $\mathcal{A}_{N-1}$, associated to this decomposition is the following important structure: the face poset, which is the set of all cells induced by $\mathcal{A}_{N-1}$ ordered by inclusion. The maximal cells (all $(N-1)$-dimensional) are called regions and are shaped as polyhedral cones, the coordinates of the points in the interior of a region obey the relation:

$$x_{i_1} < x_{i_2} < \cdots < x_{i_{N-1}} < x_{i_N} \tag{1}$$

the dominance relations (1) between the coordinates can be encoded by the sequence

$$(i_1)(i_2) \ \cdots \ (i_{N-1})(i_N) \tag{2}$$

thereafter referred as the cell dominance partition sequence (DPS), where the set of indices i_α is a permutation of $(1, 2, \ldots, N-1, N)$. Each index appears enclosed between parenthesis for reasons that will be made clear in the next section.

Reflecting a point in general position on $\mathcal{H}_{\mathbf{ij}}$ gives an image where the coordinates **i** and **j** are switched and the others are left unchanged. Multiple

reflections of a point on the hyperplanes $\mathcal{H}_{\mathbf{ij}}$ generate a set of $N!$ images, which are the permutations of its coordinates. This leads to the fact that the $\binom{N}{2}$ hyperplanes form a Coxeter reflection arrangement [8] whose symmetry group is isomorphic to the symmetric group S_N of permutations of the set $(1, 2, \ldots, N-1, N)$.

The reflection group $\mathcal{A}_{N-1}$ is also the symmetry group of a polytope: the N-permutohedron or Π_{N-1} [4], so called because its vertices are obtained by permuting the coordinates of the vector $(1, 2, \ldots, N-1, N)$. The faces of the the permutohedron are polar to the cells of the hyperplane arrangement and the face lattices of both are isomorphic.

For a molecule with N atoms as the x, y and z-coordinates are independent of each other [2] we have a $\mathcal{A}_{N-1}$ partition for each of them, that is ${\mathcal{A}_{N-1}}^3$ for the whole CS. As it has been emphasized in [2, 9] the -1 is because of the translation symmetry: the conformations outside the hyperplane $\mathcal{U}$ correspond to translated 3-D structures.

The radial dimension in CS is also spurious: multiplying the coordinates of an arbitrary 3-D conformation by a positive factor generates a set of points lying on a half-line starting at the origin. The partition $\mathcal{A}_{N-1}$ is *central* because that takes into account the scaling symmetry.

${\mathcal{A}_{N-1}}^3$ on the other hand does not take into account the rotation symmetry [9], the solution of this problem and its consequences will be discussed in sections 4–7.

3. The face lattice poset

The combinatorial structure of the $\mathcal{A}_{N-1}$ face poset is the fundamental concept behind this work, it can be understood by studying a class of objects called tournaments, which are directed graphs with $\mathcal{N}$ nodes [10], these are used to investigate the properties of permutations, so useful for characterizing the cells in CS.

A permutation of a set of $\mathcal{N}$ elements can be represented by an acyclic, complete and labelled tournament (see figure 1 for a description), where:

- The term acyclic means that the graph contains no directed cycles.
- A graph is complete if there is always an arc between any two nodes, if an arc goes from i to j we say that i dominates j. The score of a node is the number of nodes it dominates.
- Each node of the graph has a unique label, which is a number between 1 and $\mathcal{N}$ that distinguishes it from the other nodes.

In what follows the term tournament refers exclusively to tournaments where the above qualifiers apply.

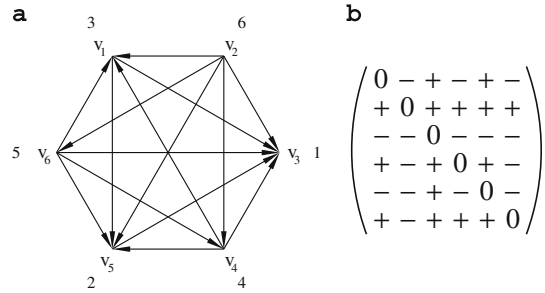

Figure 1. (a) A complete acyclic tournament corresponding to the permutation (3, 6, 1, 4, 2, 5), which is the score of each vertex plus 1, the indices in the dominance sequence of vertices (3)(5)(1)(4)(6)(2) correspond to the inverse permutation, and (b) The antisymmetric incidence matrix, the rows in the upper triangle taken in succesion form the sign vector.

For a tournament with $\mathcal{N}$ nodes the following statements are true:

I. *In a tournament there is always a node called the* sink *that is dominated by every other node*.
Consider the last node of any maximal directed path, if an arc connects it to another node then either the path is not maximal or there is a cycle; if there were a second sink it would be connected to the first and either it would dominate or be dominated.

II. *In a tournament there is always a node called the* source *that dominates every other node*.

III. *Any subgraph of a tournament is also a tournament*.
Any subgraph from a complete graph is also complete, and it can contain no cycles otherwise they would also be present in the parent graph.

IV. *There is one maximal path that spans the graph*.
Consider the subtournament obtained by removing the source, then start the path with the arc that goes from the source to the subsource, and repeat the same step with the subgraph until you reach the sink. The path obtained goes through every node since there are $\mathcal{N}-1$ steps, and is maximal since skipping a subsource for another node shortens the path since the node is dominated by the subsource.

V. *The sequence of labels of the nodes visited by the maximal path is the dominance partition sequence*.
By the construction procedure the first node, the source, dominates all other nodes, the second dominates the remaining nodes and so on.

VI. *Reversing an arc between two nonconsecutive nodes creates a directed cycle*.

Theorem 1. In a tournament the arcs between a set of consecutive nodes in the maximal path can be arbitrarily reversed and the resulting graph still be a tournament if the subgraph spanning the consecutive nodes is a tournament.

Since the subgraph and its complement are tournaments they contain no cycles, thus a cycle must involve nodes between the subgraph and the complement, but this is not possible since by construction the set of consecutive nodes is dominated by the preceeding nodes in the maximal path and likewise it dominates the following ones.

By V reversing an arc between contiguous nodes is equivalent to a transposition in the DPS.

Theorem 2. In a tournament encoded by $(i_1)(i_2)\ \cdots\ (i_\alpha)\ \cdots\ (i_{\alpha+n-1})\ \cdots\ (i_{N-1})\ (i_N)$ the permutations in the set of n consecutive indices $i_\alpha\ \cdots\ i_{\alpha+n-1}$ give a set of tournaments that encode the vertices of an n-permutohedron.

If we restrict ourselves to the n-dimensional subspace spanned by the coordinates $(x_{i_\alpha}, \ldots, x_{i_{\alpha+n-1}})$ the permutations of the indices above corresponds to the permutations of the coordinates of the vector $(\alpha, \alpha+1, \ldots, \alpha+n-1)$, which are the vertices of a Π_{n-1}.

Corollary. The n-permutohedron is a face of Π_{N-1}.

Obviously since it is contained in the affine hyperplane $x_{i_\alpha} + x_{i_{\alpha+1}} + \cdots + x_{i_{\alpha+n-1}} = n(\alpha + (n-1)/2)$. This face is encoded by the DPS

$$(i_1)(i_2)\ \ldots\ (i_\alpha \ldots\ i_{\alpha+n-1})\ \ldots\ (i_{N-1})(i_N) \tag{3}$$

that represents the set of $n!$ sequences that are permutations of the indices i_α to $i_{\alpha+n-1}$.

Corollary. The sequence $(i_1)(i_2)\ \ldots\ (i_\alpha \ldots\ i_{\alpha+n-1}) \ldots\ (i_\beta \ldots\ i_{\beta+m-1}) \ldots\ (i_{N-1})\ (i_N)$ encodes the $(n+m-2)$-face $\Pi_{n-1} \times \Pi_{m-1}$.

This can be seen from the definition given above of the product of polytopes.

Thus the meaning of parenthesis in DPSs becomes apparent: each parenthesis enclosing a sequence of length n encodes an Π_{n-1} polytope, and the whole sequence encodes the product of all these polytopes.

These sequences can be ordered by inclusion to form a face lattice poset, which is isomorph to the one obtained with the sign vectors, since like DPSs they are another encoding scheme for tournaments [2] (see figure 1).

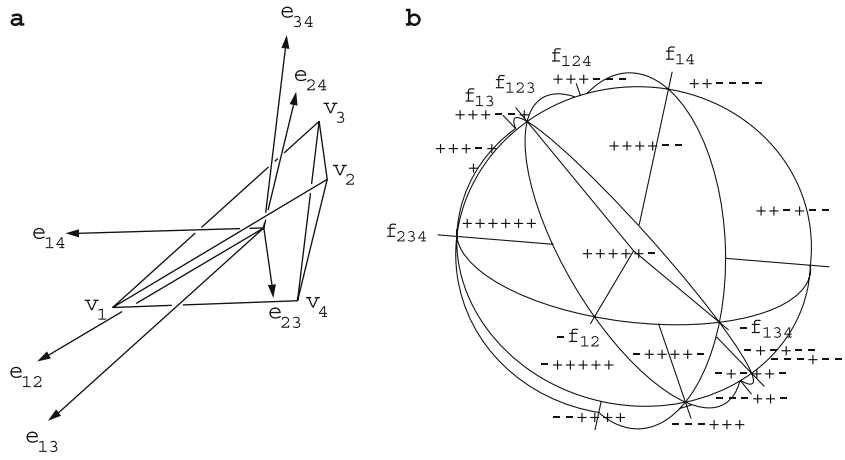

Figure 2. The $\mathcal{A}_3$ partition of a random simplex. (a) The random simplex with the vectors e_{ij} centered at the origin, and (b) The partition of 3D-space by the planes $\mathcal{E}_{ij}$ represented as intersecting disks centered at the origin, visible 3-D cells are designated by their sign vector and 1-dimensional cells are labelled by the corresponding $f_{...}$ symbols (7).

This is an important feature because it implies the modularity of the model: the face lattice of a molecule can be obtained as the product of the face lattices of subsets of atoms.

4. Enumerating the orientations of a simplex

For a simplex with random morphology we define the set of vectors that run along the edges and their associated central planes (figures 2(a) and (b))

$$e_{\mathbf{ij}} = v_{\mathbf{i}} - v_{\mathbf{j}}, \quad 1 \leqslant \mathbf{i} < \mathbf{j} \leqslant 4, \tag{4}$$

$$\mathcal{E}^0_{\mathbf{ij}}(x) = \{x \in \mathbb{R}^3 : e_{\mathbf{ij}}.\, x = 0\}. \tag{5}$$

Each plane divides 3-D space into positive and negative halves

$$\mathcal{E}^+_{\mathbf{ij}}(x) = \{x \in \mathbb{R}^3 : e_{\mathbf{ij}}.\, x > 0\} \quad \text{and} \quad \mathcal{E}^-_{\mathbf{ij}}(x) = \{x \in \mathbb{R}^3 : e_{\mathbf{ij}}.\, x < 0\}. \tag{6}$$

As for the regular tetrahedron described above (5) and (6) generate an $\mathcal{A}_3$ partition of 3-D space in 24 irregular shaped cells (figure 2(b)).

This partition has the following interesting property: assume for instance that the x-axis of a central orthogonal reference system in general position lies entirely within the cell encoded by the permutation $(3, 1, 4, 2)$, or equivalently the sign vector $(+ - + - -+)$, then the dominance relation $v_{2_x} < v_{4_x} < v_{1_x} < v_{3_x}$ holds for the x coordinates of the vertices of the simplex.

This suggests a method for enumerating cells in $\mathcal{A}_3{}^3$ that correspond to the different orientations[4] of the simplex:it suffices to enumerate the cells with the lowest dimensions, the more numerous (3, 3, 3)-dimensional cells can be easily obtained through the connecting paths in the face lattice.

The 1-dimensional cells in $\mathcal{A}_3$ are determined by the set of vectors perpendicular to the faces of the simplex and to pairs of opposite edges

$$\begin{aligned} &f_{123} = e_{12} \wedge e_{23}, \quad f_{124} = e_{12} \wedge e_{24}, \quad f_{134} = e_{13} \wedge e_{34}, \quad f_{234} = e_{23} \wedge e_{34}, \\ &f_{12} = e_{12} \wedge e_{34}, \quad f_{13} = e_{13} \wedge e_{24}, \quad f_{14} = e_{14} \wedge e_{23} \end{aligned} \tag{7}$$

their corresponding central planes will be designated $\mathcal{F}_{\mathbf{ijk}}$ and $\mathcal{F}_{\mathbf{ij}}$.
If we take the sign of the scalar products between the sets of vectors (4) and (7) we obtain a matrix

	e_{12}	e_{13}	e_{14}	e_{23}	e_{24}	e_{34}	DPS
f_{123}	0	0	+	0	+	+	(4)(123)
f_{124}	0	−	0	−	0	+	(234)(3)
f_{134}	+	0	0	−	−	0	(2)(134)
f_{234}	+	+	+	0	0	0	(234)(1)
f_{12}	0	−	−	−	−	0	(12)(34)
f_{13}	+	0	+	−	0	+	(24)(13)
f_{14}	−	−	0	0	+	+	(14)(23)

(8)

that up to a sign reversal is an invariant [3, 11], it is the same for any simplex whatever its morphology. The rows are the sign vectors of the 1-dimensional cells with the corresponding dominance partition sequence on the righ, these cells can be seen in figure 2(b), where the labels $f_{\mathbf{ijk}}$ and $f_{\mathbf{ij}}$ are on top of the lines intersected by the planes $\mathcal{E}_{\mathbf{ij}}$, $\mathcal{E}_{\mathbf{ik}}$, $\mathcal{E}_{\mathbf{jk}}$, and $\mathcal{E}_{\mathbf{ij}}$, $\mathcal{E}_{\mathbf{kl}}$, respectively.

We start by enumerating the orientations of a reference system whose z-axis is parallel to one of the vectors (7), f_{123} as an example, the remaining axis x and y will be on the plane $\mathcal{F}_{123}$, the problem is to determine how the $\mathcal{E}_{\mathbf{ij}}$s (5) divide this plane into 2-dimensional cells. In figure 3, we can see the four possible 12-sector partitions that can be generated by the vectors e_{12}, e_{13} and e_{23} and the perpendicular intersections of the planes $\mathcal{E}_{12}$, $\mathcal{E}_{13}$ and $\mathcal{E}_{23}$. This partition gives us only half of the sign vectors components, to obtain the remaining ones we need to introduce a morphological classification of simplexes.

5. Morphological classification of simplexes

For a given simplex, like the one in figure 2(a) for instance, we compute the sign of the scalar products of the vectors (4) and (7) between them, this gives the

[4] All along this work the term orientation is used interchangeably with DPS and sign vector.

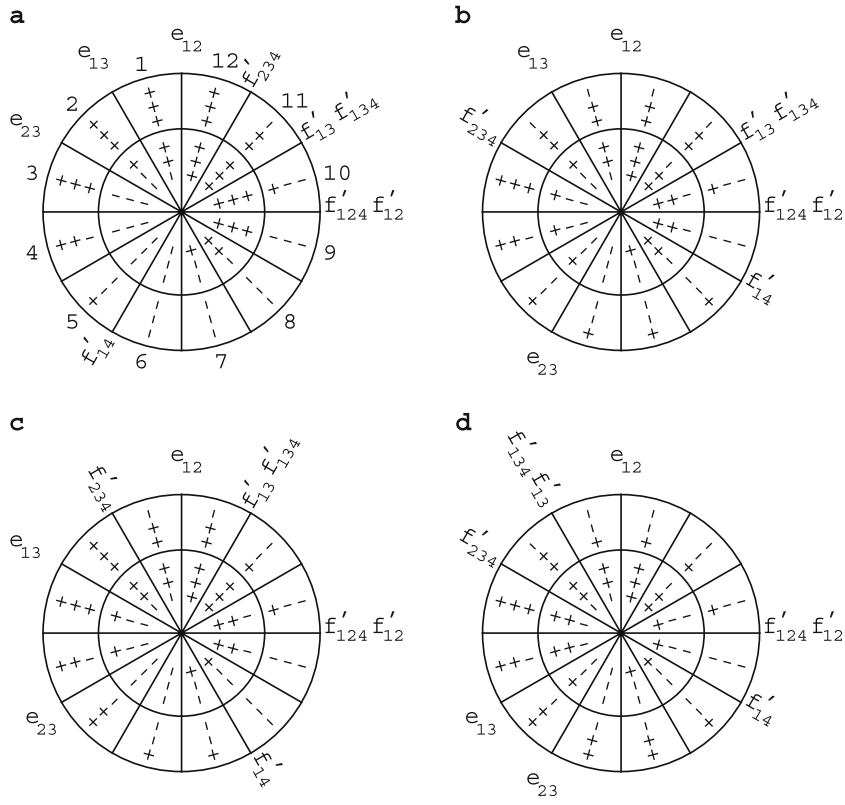

Figure 3. The four possible partitions of the plane $\mathcal{F}_{123}$. Within figures (**a**–**d**) the vector f_{123} points in the upward direction, the labels e_{12}, e_{13}, and e_{23} are over the lines that run along these vectors, and the corresponding perpendicular lines are the intersections with the planes $\mathcal{E}_{12}$, $\mathcal{E}_{13}$, and $\mathcal{E}_{23}$, respectively. The label $f'_{\mathbf{ijk}}$ means that the corresponding line runs along the projection of vector $f_{\mathbf{ijk}}$ on $\mathcal{F}_{123}$. The labels f'_{124} and f'_{12} over the intersection of plane $\mathcal{E}_{12}$, for instance, is because f_{124} and f_{12} are contained in that plane, and reciprocally e_{12} is contained in the planes $\mathcal{F}_{124}$ and $\mathcal{F}_{12}$. All these lines converge at the origin and partition $\mathcal{F}_{123}$ in 12 sectors: between the inner and outer circles are the sign vector components of e_{12}, e_{13}, and e_{23}, for each sector they should be read from inside out in that order; within the inner circle there are the sign vector components of f'_{124}, f'_{134} and f'_{234}, respectively. The sectors are numbered from 1 to 12 as indicated in **a**.

following two tables

	e_{13}	e_{14}	e_{23}	e_{24}	e_{34}
e_{12}	+	+	−	−	−
e_{13}		+	+	−	−
e_{14}			+	+	+
e_{23}				+	+
e_{24}					+

	f_{124}	f_{134}	f_{234}	f_{12}	f_{13}	f_{14}
f_{123}	+	+	+	−	+	+
f_{124}		+	+	+	+	+
f_{134}			+	+	+	−
f_{234}				−	+	−
f_{12}					+	+
f_{13}						+

. (9)

The set of signs (9) refer mostly to angles between adjacent edges and dihedral angles between contiguous faces: +, 0 and − are for acute, right and obtuse angles, respectively.

Thus the rough morphological characteristics of a simplex can be encoded in a 36 bit binary[5] sequence: there are a total of 3936 sequences that correspond to geometrically realizable simplexes, these define the set of morphological classes **A** of labelled simplexes. We define the volume of a class as the set of cells it spans in $\mathcal{A}_3{}^3$.

It should be reminded that this classification has a graph structure, since geometrical deformations in a simplex from one class induce a transition to other classes thus establishing a connectivity between them; the precise structure of such a graph is of no utility in the present work, but it is important to bear in mind this concept when dealing with the range of morphological variation of simplexes in section 8.

The binary sequence (9) is instrumental in finding the partition of the planes perpendicular to 1-dimensional cells, in our exemple it can be deduced from (9) that the partition of $\mathcal{F}_{123}$ is the one of figure 3c, since it is the only one that satisfies the relation

$$(\mathrm{SIGN}(e_{12}.e_{13}), \mathrm{SIGN}(e_{12}.e_{23}), \mathrm{SIGN}(e_{13}.e_{23})) = (+\,-\,+).$$

There are also the relations concerning vectors e_{14}, e_{24}, and e_{34}

$$(\mathrm{SIGN}(e_{14}.e_{12}), \mathrm{SIGN}(e_{14}.e_{13}), \mathrm{SIGN}(e_{14}.e_{23})) = (+\,+\,+), \tag{10a}$$
$$(\mathrm{SIGN}(e_{24}.e_{12}), \mathrm{SIGN}(e_{24}.e_{13}), \mathrm{SIGN}(e_{24}.e_{23})) = (-\,-\,+), \tag{10b}$$
$$(\mathrm{SIGN}(e_{34}.e_{12}), \mathrm{SIGN}(e_{34}.e_{13}), \mathrm{SIGN}(e_{34}.e_{23})) = (-\,-\,+) \tag{10c}$$

thus e'_{14}, the projection[6] of e_{14}, must lie in sectors 2 or 3 by (10a); similarly e'_{24} and e'_{34} must be in sectors 6 or 7 by (10b) and (10c). These ambiguities can be resolved by set of relations

$$(\mathrm{SIGN}(e_{14}.f_{123}), \mathrm{SIGN}(e_{24}.f_{123}), \mathrm{SIGN}(e_{34}.f_{123})) = (+\,+\,+), \tag{11a}$$
$$(\mathrm{SIGN}(f_{124}.f_{123}), \mathrm{SIGN}(f_{134}.f_{123}), \mathrm{SIGN}(f_{234}.f_{123})) = (+\,+\,+), \tag{11b}$$
$$(\mathrm{SIGN}(f_{12}.f_{123}), \mathrm{SIGN}(f_{13}.f_{123}), \mathrm{SIGN}(f_{14}.f_{123})) = (-\,+\,+), \tag{11c}$$

e_{14} for instance, lies on $\mathcal{F}_{124}$ and together with f_{124} stands above $\mathcal{F}_{123}$, by (11a) and (11b), this implies that $\mathrm{SIGN}(e'_{14}.f'_{124}) = -$. Repeating this procedure for

[5] We exclude sequences harboring 0s as they form a set of null measure.
[6] The $'$ superscript designates the projection of a vector on $\mathcal{F}_{\ldots}$.

f_{134} and f_{234}, and for each of the vectors e_{24} and e_{34} we end up with

$$(\mathrm{SIGN}(e'_{14}.f'_{124}), \mathrm{SIGN}(e'_{14}.f'_{134}), \mathrm{SIGN}(e'_{14}.f'_{14})) = (---), \tag{12a}$$

$$(\mathrm{SIGN}(e'_{24}.f'_{124}), \mathrm{SIGN}(e'_{24}.f'_{13}), \mathrm{SIGN}(e'_{24}.f'_{234})) = (---), \tag{12b}$$

$$(\mathrm{SIGN}(e'_{34}.f'_{12}), \mathrm{SIGN}(e'_{34}.f'_{134}), \mathrm{SIGN}(e'_{34}.f'_{234})) = (+--), \tag{12c}$$

(12a), (12b) and (12c) imply that e'_{14}, e'_{24}, and e'_{34} are to be found in sectors 3, 6 and 7, respectively, thus removing the ambiguities.

There is one ambiguity though that cannot be resolved with the binary sequence (9): $\mathcal{H}_{24}$ runs through sectors 3 and 9 together with e'_{14}, and $\mathcal{H}_{14}$ runs through sectors 6 and 12 as e'_{24} , so we end up with two possible partitions of $\mathcal{F}_{123}$ that are shown in figure 4.

As can be seen from figure 4 each partition generates 12 2-dimensional cells and the same number in one dimension, by construction the lines along the 1-dimensional cells are never perpendicular to each other, as a consequence for an (x, y) reference system centered at the origin if one of the axis runs along the edge of a sector the other will be located inside a sector: rotating the axis system enables us to scan 12 (2, 3, 1) and 12 (3, 2, 1)-dimensional cells.

Thus for any orientation structure associated with a plane $\mathcal{F}_{...}$, as those in figure 4, a reference system with one axis perpendicular to the plane can be in $2 \times 12 \times 6$ cells with dimensions any permutation of the sequence (3, 2, 1) in (x, y, z), the (3, 3, 3)-dimensional cells can be derived from these through the connecting paths in the $\mathcal{A}_3{}^3$ cell lattice poset. This solves the problem of enumerating cells that correspond to different orientations of the simplex.

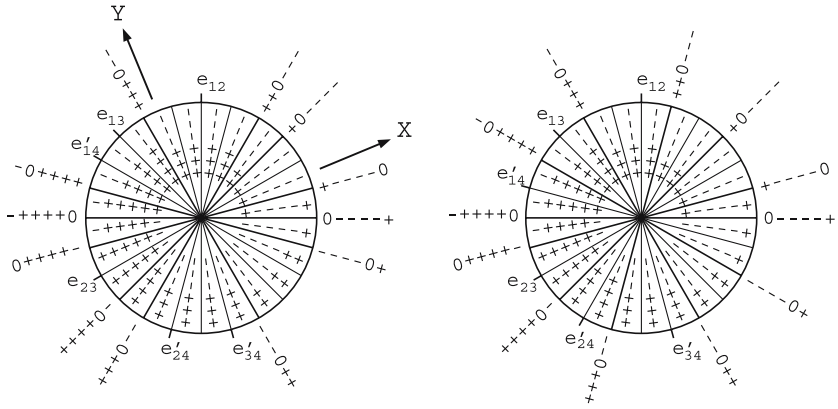

Figure 4. The two possible orientation structures of $\mathcal{F}_{123}$. The thick lines are the intersections of $\mathcal{E}_{\mathbf{ij}}$ with $\mathcal{F}_{123}$, the thin ones are lines along the vectors $e_{\mathbf{ij}}$ and $e'_{\mathbf{ij}}$. The sign vectors of the 2-dimensional cells lie inside the circle, the 1-dimensional ones are outside along the corresponding partition line, they should be read from inside out. An (X, Y) axis system has been superimposed on the first structure as a visual aid to show how the sectors can be scanned.

6. The conformational space of a simplex

We have seen that the binary sequences (9) cannot define unambiguous partitions of the planes $\mathcal{F}_{...}$: for each $\mathcal{F}_{\mathbf{ijk}}$ there can be between 1 and 3 possible orientation structures, and between 1 and 24 for each $\mathcal{F}_{\mathbf{ij}}$; in a given class only a fraction of the combinations between the different orientation structures, one from each plane, give geometrically realizable simplexes.

To remove ambiguities we need to define a set **B** of morphological classes such that for each one the range of geometrical variation allows only one orientation structure per $\mathcal{F}_{...}$. An empirical Monte Carlo calculation yields a total of 125712 classes of labelled simplexes, a class **A** has a number of subclasses **B** that goes from a minimum of 1 up to a maximum of 220. These morphological subclasses have the remarkable property that for *a given* 3-D *conformation any cell in the volume can be reached through a rotation*, which is is an obvious consequence of the one to one correspondence between $\mathcal{F}_{...}$ planes and orientation structures.

Thus a class **A** can be decomposed into a set of subclasses **B**, that can be unambiguously oriented in a standard 3-D reference frame, and its volume in CS is simply the union of the volumes of its subclasses.

7. The orientation structures

To achieve a morphological classification of simplexes we need to know how many classes of orientation structures there are, since the classes **A** decompose into subclasses **B** and each of these is determined by seven orientation structures.

A first classification concerns the circular order of the vectors $e'_{\mathbf{ij}}$ in the plane $\mathcal{F}_\alpha$. This can be deduced from the set of signs (9), for instance: by (4) and (7) the shortest circular path going through e'_{12}, e'_{13} and e'_{23} must be less than π, and it runs clockwise/counter-clockwise if the sign of $\mathcal{F}_{123}.\mathcal{F}_\alpha$ is $-/+$, respectively.

This exemple leads to the general solution that was discussed in [9]: the seven vectors (7) define a central partition dual to $\mathcal{A}_3$ [11] that divides the 3-D space in 32 cells. The sign vector of the cell that contains $\mathcal{F}_\alpha$ defines the sense of the shortest circular path that connects the projected vectors in the seven ordered sets $\{e'_{12}, e'_{13}, e'_{23}\}$, $\{e'_{12}, e'_{14}, e'_{24}\}$, $\{e'_{13}, e'_{14}, e'_{34}\}$, $\{e'_{23}, e'_{24}, e'_{34}\}$, $\{e'_{12}, e'_{34}\}$, $\{e'_{13}, e'_{24}\}$ and $\{e'_{14}, e'_{23}\}$. This generates a set of 7 constraints from which the circular order of the $e'_{\mathbf{ij}}$s in $\mathcal{F}_\alpha$ can be deduced, making a total of 32 possible circular orientations.

As can be seen in figure 4 on the plane $\mathcal{F}_\alpha$ each $e'_{\mathbf{ij}}$ contributes a total of four separations between sectors at periodic intervals of 90° each comprising exactly six sectors, on the other hand there are two classes of separations: either

a line along the vector e'_{ij} or the intersection of a plane $\mathcal{H}_{ij}$, in an interval of 90° the possible distributions of the two separators amounts to a total of 2^5 combinations. This makes 1024 classes of orientation structures like those in figure 4, among these 48 appear not to be geometrically realizable since they are not found in any class **B**.

8. Determination of the graph of cells

In mesoscopic models of biological macromolecules atoms are represented as point-like structures surrounded by an atomic force field [12, 13], thus any four atoms are the vertices of a 3-simplex. Also for a molecular system numbering its atoms from 1 to N defines an order relation that allows to designate the 3-simplexes as 4-tuples of ordered integers.

Beyond the orientation problem, the classes **A** and **B** bring the possibility of analysing the dynamics of a molecular system in terms of discrete entities, the range of morphological variation for simplexes within a molecule can be explored in MDS and the results can be summarized as follows [9, 14]:

- 90% of simplexes in a structure evolve within less than 20 classes **A**.
- The maximum variation observed is somewhat less than 200 classes, about 5% of the total.

This result opens up the possibility of determining the set of geometrically accesible cells in the CS of a molecular system.

The CS of a simplex has a total of 13824 cells and, typically, the volume of a class **A** is about one third of that number, much less if we exclude structures that can be derived through a rotation. This volume is very small when compared to the huge number of cells spanned by a molecular system, and it can be reasonably assumed that a molecular dynamics run scans the volume of a simplex. What cannot be scanned by a simulation is the set of structures that arise by combining the local movements of the molecule.

The MDSs can be used to determine the subgraph of classes spanned by every simplex, and the volume of the molecular system in CS can be obtained by progressively merging the CS of individual simplexes. As we were able to determine the different orientations of a simplex this process can be done excluding redundant rotated structures.

Before proceeding further let us show with a simple exemple the basic operations that are involved in the process of merging CSs. If we have two adjoining simplexes S_α and S_β represented by the tetrads $\{14, 33, 82, 86\}$ and $\{14, 82, 86, 91\}$, respectively (notice that their common faces correspond to the vertices (v_1, v_3, v_4) and (v_1, v_2, v_3)), if the 3-D structure of S_α is in a cell encoded

by the dominance partition sequence

$$((82)(14)(86)(33), (33)(82)(86)(14), (86)(14)(33)(82)) \tag{13}$$

then the set cells in CS_β geometrically compatible with (13) will be those whose DPS contains the pattern

$$((82)(14)(86), (82)(86)(14), (86)(14)(82)). \tag{14}$$

Thus a cell in CS_β with DPS

$$((82)(91)(14)(86), (91)(82)(86)(14), (86)(14)(91)(82)) \tag{15}$$

can be merged with (13) and generates the set of four cells in $CS_{\alpha\cup\beta}$

$$((82)(91)(14)(86)(33), (33\ 91)(82)(86)(14), (86)(14)(33\ 91)(82)), \tag{16}$$

which corresponds to a square face in the polar polytope.

To calculate the graph of the geometrically accesible cells we begin by picking an arbitrary reference simplex, preferably one with low-morphological variation, and arbitrarily choose an orientation among those available, this will be the simplex on level 1, the simplexes adjacent to this one form the level 2, and so on. Since adjacent simplexes in a 3-D structure share three vertices the shortest adjacency path between any two of them has at most length 4, so we end up with simplexes in five levels.

We need not to include every simplex from the molecule to perform a useful calculation, but there is the minimum requirement that every pair of atoms from a total of $\binom{N}{2}$ should be present at least once in a 4-tuple, otherwise the DPSs could not be determined.

The calculation can be done with the following procedure:

1. Start at level 1.
2. From any simplex in level n we select the compatible orientations in the adjoining simplexes in level $n+1$.
3. From any simplex in the level $n+1$ we select compatible orientations on the adjoining simplexes at the same level.
4. If $n < 5$ we go to step 2 and continue with level $n+1$.

A link is created between any two compatible orientations in adjacent simplexes. This is done in two steps:

1. If the simplex in the lower level has not yet been visited any orientation compatible with those from the simplex in the upper level is selected.
2. Otherwise any orientation that has not been selected is discarded. And likewise an orientation that fails to form a link with an adjacent simplex is discarded because of geometrical inconsistency.

The implementation of this procedure as an efficient computer algorithm requires that the CS of a class **A** simplex be quickly searched for orientations compatible with those from the adjoining simplexes, these can be obtained from the set of orientation structures available to each 1-dimensional cell $\mathcal{F}_{...}$ (7). This requirement can be fulfilled by building a hash table from where the DPSs like (15) can be retrived, such table has the following set of entries:

1. the number of the orientation class: from 1 to 976,
2. the connecting face, numbered from 1 to 4,
3. the 1-dimensional $\mathcal{F}_{...}$ cell (7) corresponding to the orientation structure, numbered from 1 to 7,
4. the chirality of the simplex: right or left-handed,
5. the pattern (14), of a total of 216 possible patterns.

9. Conclusion

The aim of the present work has been to bring the sheer complexity of molecular conformational space to tractable dimensions, by building a structure that encodes the set of geometrically accesible 3-D-conformations of a thermalized molecule, and putting it in a compact and manageable code. The price to pay to achieve this result is the loss of the absolute precision over the local 3-D-conformations of molecular structures [2], but in this work we only seek to obtain a global view of conformational space. The present formalism may be a useful complement to molecular dynamics simulations, that in the detailed exploration of small regions are unexcelled.

What remains to be done is to explore the graph of cells with a Hamiltonian functional over a force field and perform energy optimizations. It should be emphasized that as a Hamiltonian is a function of distances between atoms the present structure offers the possibility of calculating the energy over entire regions of CS, since the interatomic distances can be enumerated for a set of cells and in this case the energy function is nothing else than a sum over a set of coefficients.

References

[1] M. Karplus and J.A. McCammon, Molecular dynamics simulations of biomolecules, Nature Struct. Biol. 9 (2002) 949–852.
[2] J. Gabarro-Arpa, A central partition of molecular conformational space, I. Basic structures Comp. Biol. Chem. 27 (2003) 153–159.
[3] A. Bjorner, M. las Vergnas, B. Sturmfels, and N. White, *Oriented Matroids* (Cambridge University Press, Cambridge, UK, Sect. 2, 1993).
[4] S. Fomin and N. Reading, Root systems and generalized associahedra, math.CO/0505518 (2005).

[5] G. Kreweras, Sur les partitions non croisées d'un cycle, Disc. Math. 1 (1972) 333–350.
[6] C.R. Shalizi and C. Moore, What is a macrostate? Subjective observations and objective dynamics, cond-mat/0303625 (2003).
[7] P.G. Mezey, The topology of catchment regions of potential energy hypersurfaces, Theor. Chem. Acc. 102 (1999) 279–284.
[8] H.S.M. Coxeter, *Regular polytopes* (Dover Publicaions Inc., New York, 1973).
[9] J. Gabarro-Arpa, A central partition of molecular conformational space. II. Embedding 3D-structures, in: *Proceedings of the 26th Annual International Conference of the IEEE EMBS* (San Francisco, 2004) pp. 3007–3010.
[10] J. W. Moon, *Topics on Tournaments* (Holt, Rinehart and Winston, New York, 1968).
[11] J. Folkman and J. Lawrence, Oriented matroids, J. Comb. Theory B 25. (1978) 199–236.
[12] A.D. MacKerell Jr. et al., All-Atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B 102 (1998) 3586–3616.
[13] W. Wang, O. Donini, C.M. Reyes and P.A. Kollman, Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions, Ann. Rev. Biophys. Biomol. Struct. 30 (2001) 211–243.
[14] C. Laboulais, M. Ouali, M. Le Bret and J. Gabarro-Arpa, Hamming distance geometry of a protein conformational space, Proteins: Struct. Funct. Genet. 47 (2002) 169–179.