J Math Chem (2008) 44:872–883
DOI 10.1007/s10910-008-9384-5

ORIGINAL PAPER

A central partition of molecular conformational space. IV. Extracting information from the graph of cells

Jacques Gabarro-Arpa

Received: 12 November 2007 / Accepted: 22 February 2008 / Published online: 21 May 2008

Abstract In previous works (Gabarro-Arpa, J. Math. Chem. 42 (2006) 691–706) a procedure was described for dividing the $3 \times N$-dimensional conformational space of a molecular system into a number of discrete cells, this partition allowed the building of a combinatorial structure from data sampled in molecular dynamics trajectories: the graph of cells or **G**, that encodes the set of cells in conformational space that are visited by the system in its thermal wandering. Here we outline a set of procedures for extracting useful information from this structure: (1st) interesting regions in the volume occupied by the system in conformational space can be bounded by a polyhedral cone, whose faces are determined empirically from a set of relations between the coordinates of the molecule, (2nd) it is also shown that this cone can be decomposed into a set of smaller cones, (3rd) the set of cells in a cone can be encoded by a simple combinatorial sequence.

Keywords Molecular conformational space · Hyperplane arrangement · Face lattice · Molecular dynamics

1 Introduction

The aim of this series of papers [1–4] is to build a set of mathematical tools for studying the energy landscape of proteins [5–7], and the present paper is a step further towards this goal.

The energy surface of proteins is the essential tool for understanding the physico-chemistry of basic biological processes like catalysis [7]. It is also a complex

J. Gabarro-Arpa (✉)
Ecole Normale Supérieure de Cachan, LBPA, CNRS UMR 8113, 61, Avenue du Président Wilson, 94235 Cachan Cedex, France
e-mail: jga@gtran.org

multi-dimensional structure that can only be built from the knowledge of the complete dynamical history of the molecule, which is currently out of reach for conventional molecular-dynamics simulations (thereafter referred as MDS) [7]. One reason is that in an MDS trajectory the position of every atom in the molecule is calculated with an accuracy of a hundredth of angström, which quickly overhelms even the most powerful computers. The approach taken here consists in encoding the small movements of a molecular system by means of some combinatorial structure, that allows to generate the set of realizable combinations of these movements.

Within this approach, the $3D$-structures of protein molecules are encoded into binary objects called dominance partition sequences (DPS) [1–4], these are the generalization of a combinatorial structure known as noncrossing partition sequences [8]. In this context the basic structure for studying the molecular dynamics is the set of $3D$-conformations that have the same DPS, these form a connected region in molecular conformational space[1] (in what follows abridged to *CS*) called cell, thus DPSs generate a partition of *CS* into disjoint cells. Partitions are a useful tool for studying multi-dimensional spaces, in our case they systematically span a much wider volume range than the set of points along a random trajectory curve generated by a MDS, they have also been used in many other contexts [5,6,9].

The aim of the preceding papers [1–4] was to construct a graph whose nodes are the cells visited by the molecular system in its thermal wandering, two important properties of partition sequences make this construction possible:

1. DPSs are hierarchical structures: partition sequences encoding different sets of cells can be merged into a new partition sequence encoding the union set, and the process can be repeated with the new sets of cells, thus creating a hierarchy. The importance of this property is that climbing the hierarchy ladder *the number of cells increases exponentially while the sequence length increases only linearly*. This compact coding makes possible the construction of a graph representing huge regions of *CS* whose size does not exceed the memory of a workstation computer, while keeping at the same time the essential information about the molecular structures.
2. DPSs are modular structures: partition sequences can be decomposed into subsequences that are embedded in different conformational subspaces. This allows to define a composition law: if two partition sequences from two different subspaces share the same sequence for the intersection subspace, then joining both sequences gives a realizable sequence [4].[2]

The first property tells us that the graph can be constructed, the second suggests how to build it: a molecular structure can be decomposed into sets of four atoms, its smallest $3D$ components, by composing the graphs of these one can build the graph of the molecule.

Atoms in MDSs are represented as pointlike structures surrounded by a force field [10,11], the convex envelope of a set of 4 points in $3D$-space is an irregular polytope

[1] For an N-atom molecule it is a $3 \times N$-dimensional space where each point corresponds to a $3D$ molecular conformation.

[2] That corresponds to an existing set of cells.

called a 4-simplex or simplex.[3] The conformational space of these sets is relatively small with 13,824 cells, of these only a fraction is visited by the system. With a *CS* so small it can be plausibly assumed that the accessible cells are all visited during a MDS run.

The method for building the graph that was proposed in [2] consists in

1. Establishing a morphological classification of simplexes, where each class is defined by a set of geometrical constraints.
2. The geometrical constraints that define a class allow to calculate the set of accessible cells in a simplex *CS* [4], thus to each class we can associate a graph where the nodes are the cells from this set with edges towards adjacent cells.
3. On the other hand computer simulations of protein dynamics show [2,4] that in a protein structure the majority of simplexes evolve within a reduced number of morphologies. For each 4-atom set in the molecule the graph of its *CS* is built by merging the graphs of the visited simplex morphologies.
4. The *CS* graph of the molecule, that was called the graph of cells or **G** in [4], can be built by composing the *CS* graphs of the different simplexes.

The graph of cells allows to enumerate exactly the set of visited cells in conformational space, but since the cells are encoded in a compact form unwrapping them completely is probably algorithmically hopeless. Instead here we propose the construction of more manageable coarse-grained encodings that, using the information from **G**, can be recursively decomposed into progressively fine-grained ones. This subject is developed in the next five sections:

- Section 2 is a graph of cells oriented description of the basic mathematical framework.
- Section 3 is about the basic mathematical properties of **G**.
- Section 4 describes how to determine, from empirical data, a conical boundary for the region occupied by the system in *CS*.
- Section 5 shows how to decompose this cone boundary into a set of smaller cones.
- Section 6 is devoted to describing a combinatorial sequence that encodes the conical boundary in its most compact form.

2 The basic construction

It was shown [1] that the conformational space of a molecule of $N+1$ atoms $\mathbb{R}^{3\times N}$[4] could be described to a fair degree of accuracy by means of the partition generated by a set of hyperplanes passing through the origin that form a Coxeter reflection arrangement[5] denominated $\mathcal{A}^N$ [8,12], moreover the reflections form a symmetry group that is isomorphic to the symmetric group.

In our description of *CS* we have three independent arrangements one for each coordinate (x, y, z), i.e. $\mathcal{A}^{3\times N} = \mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^N$, that generate three partitions of

[3] In what follows this denomination will be used to designate ordered sets of 4-atoms/points.

[4] $N+1$ is because the translation symmetry makes one dimension spurious [1,4].

[5] So called because a reflexion through one of the hyperplanes leaves the arrangement unchanged.

$\mathbb{R}^{3\times N}$, each dividing $\mathbb{R}^N$ into a hierarchical set of regions shaped as polyhedral cones denominated cells. The hyperplanes in our partition are defined as

$$\mathcal{H}_{ij} : x_i - x_j = 0, \quad 1 \leq i < j \leq N+1 \tag{1}$$

each $\mathcal{H}_{ij}$ divides $\mathbb{R}^N$ into three regions:

$$x_i < x_j, \quad x_i = x_j \quad \text{and} \quad x_i > x_j \tag{2}$$

in the first case we say that x_j dominates x_i, in the second case neither x_i nor x_j dominates, in the last case x_i dominates x_j. As cells are bounded by the hyperplanes (1) a consequence of (2) is that the points inside a given cell (in x, y or z) have the following property:

$$x_{i_1} \leq x_{i_2} \leq x_{i_3} \leq \cdots \leq x_{i_{N-2}} \leq x_{i_{N-1}} \leq x_{i_N} \tag{3}$$

where the sequence $(i_1, i_2, i_3, \ldots i_{N-2}, i_{N-1}, i_N)$ is a permutation of the set $\mathcal{Z}_{N+1} = (1, 2, 3, \ldots\ N, N+1)$, reflecting a point through $\mathcal{H}_{ij}$ is equivalent to permute the coordinates i and j [8]. Thus a cell where a strict "less than" relation holds for every pair of coordinates in (3) is encoded by the dominance sequence

$$(i_1)(i_2)(i_3)\ldots(i_{N-2})(i_{N-1})(i_N) \tag{4a}$$

while for a cell where $x_{i_\alpha} = x_{i_{\alpha+1}} = \cdots = x_{i_{\alpha+r}}$, for $r+1$ consecutive indices $(i_\alpha, i_{\alpha+1}, \ldots i_{\alpha+r})$ in (3) will be encoded by the dominance sequence

$$(i_1)(i_2)(i_3)\ldots(i_\alpha i_{\alpha+1}\ldots i_{\alpha+r})\ldots(i_{N-1})(i_N)(i_{N+1}) \tag{4b}$$

the first (4a) represents an N-dimensional cell while (4b) is a $(N-r)$-dimensional cell because it corresponds to the intersection of the hyperplanes $\mathcal{H}_{ij}$ with $i, j \in (i_\alpha i_{\alpha+1}\ldots i_{\alpha+r})$.

Definition 1 The position of a coordinate x_i^c in a cell of dimension N is the position of the index i in the dominance sequence of c.

An alternative encoding of cells is by means of an $N \times N$ antisymmetric sign matrix $\mathcal{S}^c$, where c stands for x, y or z. Let $1 \leq i < j \leq N+1$, then for an arbitrary point x the matrix elements $\mathcal{S}^c$ for the c coordinates are defined:

$$\begin{aligned} \mathcal{S}^c_{ij} &= - \ \text{ if } \ x_i^c < x_j^c \\ \mathcal{S}^c_{ij} &= 0 \ \text{ if } \ x_i^c = x_j^c \\ \mathcal{S}^c_{ij} &= + \ \text{ if } \ x_i^c > x_j^c \end{aligned} \tag{5}$$

As it was explained in [1,4] a direct consequence of (3) is that $\mathcal{S}^c$ can be interpreted as the incidence matrix of a digraph with no directed cycles, and the cell encodings (3) and (5) can be readily interconverted into one another

Lemma 1 *Contiguous cells in space have different dimensionalities.*

Crossing to a contiguous cell implies going between two regions in (2), so one element $\mathcal{S}_{ij}^{c}$ in (5) changes its value, and this change can never be between + and − because this would mean crossing $\mathcal{H}_{ij}^{c}$ avoiding the region $c_i = c_j$.

Definition 2 A contiguous set are all the n-dimensional cells contiguous to a $(n-1)$-dimensional separator cell.

This allows to build a hierarchical structure: the cell lattice poset, that results from ordering contiguous cells by dimensionality [1,13].

Consider two arbitrary subpartitions $\mathcal{A}_a^{d_a}$ and $\mathcal{A}_b^{d_b}$ of $\mathcal{A}^N$, corresponding to the sets of indices $\chi_a = (i_{a_1}, i_{a_2}, \ldots i_{a_{d_a+1}}) \subset \mathcal{Z}^{d_a+1}$ and $\chi_b = (i_{b_1}, i_{b_2}, \ldots i_{b_{d_b+1}}) \subset \mathcal{Z}^{d_b+1}$, respectively, and let $\chi_{a\cap b} = \chi_a \cap \chi_b$ be the set of indices that are common to both partitions.

Definition 3 Two cells $\zeta_a \in \mathcal{A}_a^{d_a}$ and $\zeta_b \in \mathcal{A}_b^{d_b}$ with sign matrices $\mathcal{S}^a$ and $\mathcal{S}^b$, respectively, are said to be compatible if $\mathcal{S}_{ij}^{a} = \mathcal{S}_{ij}^{b} \;\forall\, i, j \in \chi_{a\cap b}$.

Lemma 2 *The cell $\zeta_a \in \mathcal{A}_a^{d_a}$ is the projection of all the cells in $\mathcal{A}^N$ whose sign matrix $\mathcal{S}$ is such that $\mathcal{S}_{ij} = \mathcal{S}_{ij}^{a} \;\forall\, i, j \in \chi_a$.*

This is an immediate consequence of (3) and (5).

Let Ξ_a and Ξ_b be the set of cells in $\mathcal{A}^N$ that are projected on ζ_a and ζ_b, respectively

Lemma 3 *The set $\Xi_a \cap \Xi_b$ is non empty iff ζ_a and ζ_b are compatible.*

Suppose we have $\xi \in \Xi_a$ but $\xi \notin \Xi_b$, this means that the relative positions of the set of indices $\chi_{b\setminus a} = \chi_b \setminus \chi_a$ in the dominance sequence (4) is not the same as in ζ_b, since the reflexion group of the arrangement is the symmetric group there always will be a set of permutations/reflections that sorts the indices $\chi_{b\setminus a}$ in the dominance sequence in the same order as in ζ_b, this generates a cell $\xi' \in \Xi_a \cap \Xi_b$.

3 The graph of cells

Lemmas 2 and 3 suggest that $\mathcal{A}^{3\times N}$ can be built by merging partitions of lower dimensionality. The smallest $3D$ system is a set of 4 atoms, and $\mathcal{A}^{3\times 4-1}$, the partition of its *CS*, has exactly 13,824 cells, a computational complexity within the range of a desktop computer. Moreover, as stated in the introduction it can be reasonably assumed that such small *CS* can be thoroughly scanned by a MDS.

Following the procedure proposed in Refs. [2–4] (outlined in the introduction) we can build the *CS* of a molecular system from the *CS* of the simplexes. For this, we need to construct the graph of cells or **G** which is defined as follows:

Definition 4 Two simplexes are adjacent if they share a face.

Definition 5 The nodes of **G** are the visited cells of each simplex with edges towards the compatible cells of adjacent simplexes.

Definition 6 A transversal is a subgraph of **G** with nodes exactly one cell from every simplex such that every two cells from adjacent simplexes are compatible.

G embodies all the information contained in the *CS* of a molecular system since

Theorem 1 *The cells in a transversal are the projections of a single cell in CS*

By Lemma 3 the cells in the transversal are the projection of at least one cell in $\mathcal{A}^{3\times N}$, that cell is unique because if there were two, for instance, their sign matrices would not be the same, say that the element $\mathcal{S}^c_{ij}$ is different, then there is a set of $\binom{N-1}{2}$ simplexes that harbor the indices i and j and within this set each simplex is adjacent to $2 \times (N-3)$ other simplexes, from Definition 5 adjacent simplexes have to be compatible and the element ij in their sign matrix must be the same for all, invalidating our assumption.

Corollary *In* **G** *a node that fails to form an edge with an adjacent simplex cannot exist since it is geometrically inconsistent.*

A useful structure derived from **G** is its compact form **C** obtained by recursively substituting every contiguous set of n-dimensional nodes by their $(n-1)$-dimensional separator cell.

Finally a cell from $\mathcal{A}^{3\times N}$ is a class in an equivalence relation, since it contains all the $3D$-structures that have the same dominance sequence. In what follows we use the terms cell and $3D$-structure interchangeably.

4 Determining a conical boundary for the molecular dynamics trajectory

G is a huge structure and it is probably useless to try to explore it in full, rather the approach we take here is how to focus on regions (subgraphs) where we can expect to extract useful information. We start with the problem of finding the bounds of interesting regions, with a concrete example concerning a 2.1 ns pancreatic trypsin inhibitor (PTI) [14] MDS that was fully described in [15].

As in [15] we restrict ourselves to study the motion of C^n_α carbons each bearing a number n that reflects the linear order of residues along the polypeptide chain, as our description of *CS* is strictly modular any conclusion that can be drawn on any subset of atoms is automatically valid for the whole structure.

An information easily extracted from a MDS are the dominance relations matrices DR^c, where c stands for either x, y or z, each element of these matrices defines the equation of a face in a polyhedral cone, it encloses the region that the molecular system occupies in *CS*. The determination of the DR^cs from the MDS [15] takes the following steps:

- First, the simplex corresponding to the residue numbers $\mathsf{S}_r = \{6, 36, 40, 47\}$ was selected as the reference simplex because all along the MDS it stays within one morphological class, and because it spans a wide volume across the molecule.
- Second, the coordinates of S_r in the 1^{st} MD frame were taken as a reference and the other frames were rotated and translated so that the RMS between $\mathsf{S}_r(1)$ and $\mathsf{S}_r(f)$ be a minimum [16].

- Third, the quantities DR^c_{ij}, $1 \leq i < j \leq N+1$, were determined
  - $DR^c_{ij} = +$, $DR^c_{ji} = -$ if $C^i_{\alpha_c} > C^j_{\alpha_c}$ for all coordinate frames.
  - $DR^c_{ij} = -$, $DR^c_{ji} = +$ if $C^i_{\alpha_c} < C^j_{\alpha_c}$ for all coordinate frames.
  - $DR^c_{ij} = DR^c_{ji} = 0$ if neither of the above relations holds. Also, by convention $DR^c_{ii} = 0$.

The meaning of the matrix elements is obvious, if $DR^c_{ij} = +/-$ the trajectory always stays on the positive/negative side of $\mathcal{H}^c_{ij}$ (2), for $DR^c_{ij} = 0$ the trajectory can be on either side of $\mathcal{H}^c_{ij}$. The matrices for x, y and z for the MDS [15] are shown in Fig. 1, the number of non-zero terms in the matrix is the dimension of the cone.

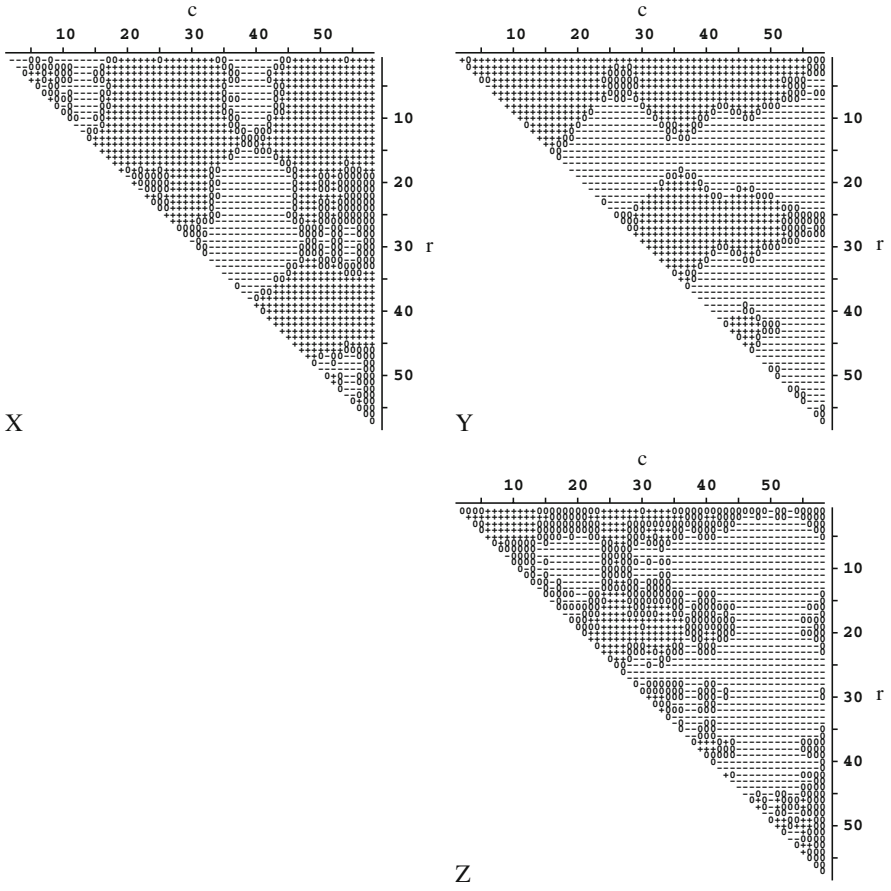

Fig. 1 Antisymmetric dominance relations matrices for the C_α coordinates, only the upper triangle is shown. For sake of clarity row and column amino acid numbers can be read from the annotated axes r and c. A matrix element can have three values: $+$, $x_r > x_c$ for all coordinate frames in the molecular dynamics run; $-$, $x_r < x_c$ for all coordinate frames in the molecular dynamics run; 0, neither of the above relations holds

Lemma 4 *The minimum position* $\min_\mu^c$ *of a coordinate* c_μ *is the number of matrix elements* $DR_{\mu j}^c = +$ *plus* $1,\ 1 \le j \le N,\ j \ne \mu$, *and the maximum position* $\max_\mu^c$ *is the minimum position plus the number of matrix elements* $DR_{\mu j}^c = 0,\ 1 \le j \le N$.

5 The fragmentation of the cone

The dominance relations matrices DR^c encode a lot of information about the structure of the volume occupied by the system in *CS*. They give us the range of positions of a given coordinate in the dominance sequence (3).

The index μ in the dominance sequence must always stay to the right of the elements it dominates if there are n_+ of such elements the minimum position of μ is $n_+ + 1$, on the other hand be n_0 the number of indifferent relations, μ can be either to the right or to the left of any of these then the maximum position of μ must be $n_+ + n_0 + 1$.

We can also extract from DR^c sets of lower dimensional cells, these are useful for fragmenting **G** into subgraphs of more manageable size. To do this we can proceed as follows: we select indices μ and ν such that

$$\begin{aligned} \forall c \in \{x, y, z\} :\ & \max_\mu^c > \min_\nu^c, \quad \max_\nu^c > \min_\mu^c \qquad \text{and} \\ & \mathrm{MIN}(\max_\mu^c, \max_\nu^c) - \mathrm{MAX}(\min_\mu^c, \min_\nu^c) \ge h^c \end{aligned} \tag{6}$$

with $h^c = 1,\ 2,\ 1$ for $DR_{\mu\nu}^c = -1,\ 0,\ 1$, respectively.

We thus select pairs of atomic indices μ and ν whose ranges overlap in x, y and z simultaneously with intersection length $\ge (h^x, h^y, h^z)$, respectively. For every pair index their ranges in any dimension are divided into three segments: left, middle (the intersection) and right; μ, for instance, can occupy any position in the left and middle segments, while ν can be in the middle and right ones, this makes a total of 3 possibilities, 4 if $DR^c = 0$ in which case μ and ν can be simultaneously in the middle segment. Obviously this can be extended to more than 2 indices: if $\mu\nu$, $\mu\omega$ and $\nu\omega$ have overlapping ranges, for instance, then there is a common overlapping range for μ, ν and ω too, which in turn gives segmentation and occupation patterns for $\mu\nu\omega$.

The importance of overlapping indices is twofold:

1. A set of molecular conformational states can be determined from them using a minimum number of cells from **G**: the indices being the same for x, y and z makes that occupation patterns for overlapping μ and ν, for instance, can be deduced from the cells in **G** corresponding to the simplexes that bear these indices.
2. One can address the basic problem of how occupation states in the dominance sequence are correlated between different coordinates.

The set of allowed overlapping indices than can be deduced from the DR^c matrices in Fig. 1 is given in Table 1.

This allows a procedure for fragmenting the cone in Fig. 1 into smaller ones. From **G** we can deduce for each set of indices from Table 1 a number of local conformations, the valid combinations of these conformations will give us smaller cones whose cells have *DPS*s with mean positions (x, y, z) much closer to the values of the cells in **G**.

Table 1 The complete sets of indices for the α-carbons of the MDS described in [15] that conform to (6)

```
( 1  2) ( 1  6) ( 1 56) ( 1 57) ( 1 58)
( 2  3) ( 2  4) ( 2  5) ( 3  4) ( 3  5)
( 3  6) ( 4  5) ( 4  6) ( 5  6) ( 6  7)
( 6 25) (10 11) (11 12) (11 35) (12 13)
(12 39) (13 14) (14 15) (15 16) (16 17)
(17 18) (18 19) (19 34) (20 46) (21 32)
(21 45) (23 24) (24 25) (24 26) (24 27)
(24 28) (25 26) (25 27) (25 28) (26 27)
(26 28) (27 28) (27 29) (28 29) (28 57)
(28 58) (29 57) (29 58) (33 34) (36 37)
(37 38) (38 39) (39 40) (41 42) (44 45)
(46 47) (48 49) (49 50) (52 53) (52 55)
(52 58) (53 54) (53 55) (53 56) (53 57)
(53 58) (54 55) (54 56) (54 57) (54 58)
(55 56) (55 57) (55 58) (56 57) (56 58) (57 58)
( 1 56 57) ( 1 56 58) ( 1 57 58) ( 2  3  4) ( 2  3  5)
( 2  4  5) ( 3  4  5) ( 3  4  6) ( 3  5  6) ( 4  5  6)
(24 25 26) (24 25 27) (24 25 28) (24 26 27) (24 26 28)
(24 27 28) (25 26 27) (25 26 28) (25 27 28) (26 27 28)
(27 28 29) (28 29 57) (28 29 58) (28 57 58) (29 57 58)
(52 53 55) (52 53 58) (52 55 58) (53 54 55) (53 54 56)
(53 54 57) (53 54 58) (53 55 56) (53 55 57) (53 55 58)
(53 56 57) (53 56 58) (53 57 58) (54 55 56) (54 55 57)
(54 55 58) (54 56 57) (54 56 58) (54 57 58) (55 56 57)
(55 56 58) (55 57 58) (56 57 58)
( 1 56 57 58) ( 2  3  4  5) ( 3  4  5  6) (24 25 26 27) (24 25 26 28)
(24 25 27 28) (24 26 27 28) (25 26 27 28) (28 29 57 58) (52 53 55 58)
(53 54 55 56) (53 54 55 57) (53 54 55 58) (53 54 56 57) (53 54 56 58)
(53 54 57 58) (53 55 56 57) (53 55 56 58) (53 55 57 58) (53 56 57 58)
(54 55 56 57) (54 55 56 58) (54 55 57 58) (54 56 57 58) (55 56 57 58)
(24 25 26 27 28) (53 54 55 56 57) (53 54 55 56 58) (53 54 55 57 58) (53 54 56 57 58)
(53 55 56 57 58) (54 55 56 57 58)
.(53 54 55 56 57 58)
```

6 A combinatorial sequence for encoding cones

The codification of cones in conformational space could be much simplified by introducing a simple extension in the formalism used for encoding dominance sequences: we allow expressions enclosed between parenthesis to overlap, and we distinguish between pairs of enclosing parenthesis by numbering them. Let us assume, for example, that we have a cone in *CS* with *DR* matrix

$$\begin{array}{cccccccc} & 1 & 3 & 4 & 7 & 8 & 9 \\ 1 & & + & + & 0 & 0 & 0 \\ 3 & - & & 0 & - & 0 & 0 \\ 4 & - & 0 & & - & 0 & 0 \\ 7 & 0 & + & + & & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & \end{array}. \tag{7}$$

The sequence

$$\overset{1}{(}3\ 4\ \overset{2}{(}8\ 9\overset{1}{)}\ 1\ 7\overset{2}{)} \tag{8}$$

is meant to encode in one formula the sequences (3 4 8 9)(1 7) and (3 4)(1 7 8 9), these represent the totality of cells from *CS* that lie inside the cone (7); structures like (8) will be designated as: generalized compact dominance sequences (*GCDS*). Notice

that parenthesis enclosed within parenthesis are not allowed within *GCDS*s since they are meaningless as dominance sequences.

*GCDS*s can encode huge numbers of cells from *CS*, for instance the first ten α-carbons in our structure [14,15] evolve within a cone in $\mathcal{A}^{3\times 10}$ encoded by the formula

$$\begin{array}{l}\overset{1}{\{}\overset{1}{\{}\overset{2}{(}1\ \overset{2}{(}5\ 8\ \overset{3}{(}6\overset{3}{)}\ \overset{1}{(}2\ \overset{4}{(}9\overset{4}{)}\ 7\overset{1}{)}\ \overset{5}{(}10\overset{5}{)}\ 3\ 4\overset{2}{)}\overset{5}{)}\}_x,\\ \overset{1}{\{}\overset{1}{(}10\overset{1}{)}\ \overset{1}{(}9\overset{1}{)}\ \overset{2}{(}8\overset{2}{)}\ \overset{3}{(}7\overset{3}{)}\ \overset{4}{(}5\ \overset{5}{(}4\overset{5}{)}\ 6\overset{4}{)}\ \overset{6}{(}2\ \overset{7}{(}3\overset{7}{)}\ 1\overset{6}{)}\}_y,\\ \overset{1}{\{}\overset{1}{(}8\ \overset{2}{(}7\ 10\overset{2}{)}\overset{1}{)}\ \overset{3}{(}6\ 9\overset{3}{)}\ \overset{4}{(}3\ 4\ 5\ \overset{6}{(}1\overset{6}{)}\ 2\overset{4}{)}\}_z\}\end{array} \tag{9}$$

that can be easily checked by comparing it with the 10×10 upper-left submatrices in Fig. 1.

Not all the cones in *CS* can be represented by *GCDS*s. A simple example will show us that the x-component of (9) cannot be extended beyond the 14th C_α. Let

	2	3	4	10	12	15
2		−	−	0	−	−
3	+		0	0	⊖	0
4	+	0		0	⊖	0.
10	0	0	0		0	0
12	+	⊕	⊕	0		0
15	+	0	0	0	0	

(10)

be the DR^x submatrix of the α-carbons 2, 3, 4, 10, 12 and 15, it gives the sequence

$$\overset{1}{(}2\ \overset{2}{(}10\overset{2}{)}\ \overset{1}{(}3\ 4\ 15\overset{1}{)}\ \overset{3}{12}\overset{3}{)} \tag{11}$$

which is clearly inconsistent because 12 dominates 3 and 4 but is on the same dominance level with 10. One can perform slight modifications in (10) that transform (11) into a valid compact formula: setting to zero the circled components in (10) gives the DR^x matrix of a *GCDS*-cone that encloses the cone (10). This modification allows us to extend the generalized dominance sequence x-component of (9) to the 20 α-carbons

$$\begin{array}{l}\overset{1}{\{}\overset{1}{(}19\overset{1}{)}\ \overset{2}{(}20\ \overset{3}{(}18\overset{3}{)}\ \overset{2}{(}1\overset{2}{)}\ \overset{4}{(}17\ \overset{5}{(}5\ \overset{6}{(}8\ \overset{7}{(}6\overset{7}{)}\ \overset{4}{(}2\ \overset{8}{(}9\overset{8}{)}\ 11\ 16\overset{4}{)}\overset{6}{)}\ \overset{7}{(}10\overset{7}{)}\ \overset{9}{(}3\ 4\overset{9}{)}\ \overset{10}{15}\overset{10}{)}\ \overset{11}{(}12\overset{11}{)}\ \overset{12}{(}14\overset{12}{)}\ \overset{5}{13}\overset{5}{)}\}_x\end{array} \tag{12}$$

One can easily verify for every coordinate from Fig. 1 that the cone that bounds the evolution in *CS* of every 10 consecutive residues of the molecular structure is a *GCDS*-cone, for chains about 20 residues and more this is generally no longer possible unless the value of some DR^c elements are made zero as in (10). This result would seem to suggest that in the MDS from [15] thermodynamic equilibrium has not been attained, for instance in (10) $C^x_{\alpha 10}$ can swap dominance with $C^x_{\alpha 3}$, $C^x_{\alpha 4}$, $C^x_{\alpha 7}$ and $C^x_{\alpha 15}$, but pairs of cells with conformations where $C^x_{\alpha 3}$, $C^x_{\alpha 4}$ and $C^x_{\alpha 7}$ cross one another on the x-axis have not been visited by the MDS. This example clearly shows that *GCDS*-cones not only

have a simple elegant formula to describe them but also they maximize the number of available states (i.e. entropy), both properties make them very convenient tools for studying *CS*.

By setting to zero a minimum number of DR^cs: 27 in x, 6 in y and 91 in z (1.9%, 0.4% and 6.6%, respectively). We obtain a *GCDS*-cone for the whole α-carbon chain

$$
\begin{aligned}
&\{\{(49\,(48\,(29\,(27\,28\,30\,(31)\,52)\,47\,(32\,(53)\,50)\,(26)\,51)\,21\,23\,24\,(19\,(20\,(25\,33)\\
&46\,55\,(54)\,22)18\,34)\,45)\,(17)\,(5\,44\,(8\,(6)\,35\,43\,(9)\,16\,(11)\,7)\,36\,(3\,4\\
&(10)\,42\,(37)\,(15)\,(12)\,(41)\,40)\,38\,(14)\,13)\,39)\}_x,\\
&\{(15\,16\,(17)\,(14)\,(18)\,(36\,(13)\,37)\,(19\,(34)\,12\,35\,38)\,(\,11)\,20\,33\,(39)\,(\,46\,(10)\\
&32\,40\,47)\,(21)\,(45)\,44\,(31\,(9\,48)\,41)\,(\,22)\,(42\,49\,50)\,8\,30\,43\,51)\,(23\\
&(\,24)\,(7\,52\,(29\,(4\,53\,54)\,(26\,27)\,5)\,6\,25\,28\,55)\,3)\}_y,\\
&\{(26\,(27\,(8\,10\,(7\,(25\,(11\,(13)\,9)\,6\,24\,(12\,(28)\,33)\,31)\,(34\,(15\,)\,32\,(29\,(17)\\
&(14)\,5\,23\,(4\,22\,35\,36\,40\,41\,(3)\,30)\,39)\,16\,21\,43)\,(38)\,18\,19)\,20)\\
&37\,42\,44)\,(55)\,(45\,48\,(52\,)\,51)\,(46\,47)\,54\,(49\,53)\,50)\}_z\}
\end{aligned}
\tag{13}
$$

This sequence sets the boundary for the molecular dynamics trajectory in [15] in a compact form.[6]

7 Conclusion

This paper is an outline of a methodology for the exploration of *CS*.

In [1–4] it was assumed that the small local movements of a molecule can be thoroughly sampled in a MDS, and a procedure was devised for building the whole set of structures that result from the combinations of these small movements. The result is a combinatorial structure called the graph of cells, that gives a global view of a molecular system dynamical conformations.

Although the graph of cells can be fitted in a desktop computer file it encodes a huge amount of structures, the present paper is a first step in solving the problem of managing this great quantity of information. Three issues have been addressed:

1. we can give bounds that delimit interesting regions (cones) in *CS*,
2. these cones can be decomposed into a set of smaller ones,
3. it is shown that cones in *CS* can be described by a combinatorial sequence

This last structure, the generalized compact dominance sequence, has embedded in it the whole set of dominance sequences that are in a cone, and can be hierarchically

[6] α-carbons from end-residues 1, 2, 56, 57 and 58 are not included because they add disorder, unnecessarily augmenting the volume of the cone without adding much information.

decomposed into a poset structure. On the other hand the graph of cells can be seen as a set of constraints between the x, y and z components of the allowed dominance sequences, then the *GCDS*s and the graph of cells complement each other beautifully, since the conformations of the molecular system can obtained by pruning the poset structure from the *GCDS* with the constraints from the graph of cells. Moreover, *GCDS*s also have a graphical structure where paths and graphical distances between cells (or $3D$-structures) can be determined, and graphical distances between atoms in a $3D$-structure can be enumerated as well. That makes *GCDS*s well suited as the base structures for the development of a combinatorial Hamiltonian in conformational space.

These issues will be further explored in forthcoming works of this series.

References

1. J. Gabarro-Arpa, A central partition of molecular conformational space. I. Basic structures, Comp. Biol. Chem. **27**, 153–159 (2003)
2. 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, pp. 3007–3010, 2004
3. J. Gabarro-Arpa, Combinatorial determination of the volume spanned by a molecular system in conformational space, Lect. Series Comput. Comput. Sci. **4**, 1778–1781 (2005).
4. J. Gabarro-Arpa, A central partition of molecular conformational space. III. Combinatorial determination of the volume spanned by a molecular system in conformational space, J. Math. Chem. **42**, 691–706 (2006).
5. P.G. Mezey, *Potential energy hypersurfaces* (Elsevier, Amsterdam, 1987)
6. D.J. Wales, *Energy landscapes* (Cambridge University Press, Cambridge, 2003)
7. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins, Nature **450**, 964–972 (2007)
8. S. Fomin, N. Reading, Root systems and generalized associahedra, math.CO/0505518 (2005)
9. C.R. Shalizi, C. Moore, What is a macrostate? Subjective observations and objective dynamics, cond-mat/0303625 (2003)
10. A.D. MacKerell Jr., et al., All-Atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B **102**, 3586–3616 (1998)
11. W. Wang, O. Donini, C.M. Reyes, P.A. Kollman, Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions, Annu. Rev. Biophys. Biomol. Struct. **30**, 211–243 (2001).
12. H.S.M. Coxeter, *Regular polytopes* (Dover Publications Inc., New York, 1973)
13. A. Bjorner, M. las Vergnas, B. Sturmfels, N. White, *Oriented matroids* (Cambridge University Press, Cambridge, UK, sect. 2, 1993)
14. M. Marquart, J. Walter, J. Deisenhofer, W. Bode, R. Huber, The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors, Acta Crystallogr. Sect. B **39**, 480–490 (1983)
15. J. Gabarro-Arpa, R. Revilla, Clustering of a molecular dynamics trajectory with a Hamming distance, Comp. Chem. **24**, 693–698 (2000)
16. W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors, Acta Cryst. **A34**, 827–828 (1978)