

Topology-based stereochemistry representation

Andreas Dietz, Christophe Fiorio, Michel Habib, Claude Lorenço

► **To cite this version:**

Andreas Dietz, Christophe Fiorio, Michel Habib, Claude Lorenço. Topology-based stereochemistry representation. Comptes rendus de l'Académie des sciences. Série IIc, Chimie, Elsevier, 1998, 1 (II), pp.95-100. 10.1016/S1251-8069(97)86267-7. lirmm-01167951

HAL Id: lirmm-01167951

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-01167951>

Submitted on 25 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Topology-based stereochemistry representation

Andreas DIETZ ^{a, b*}, Christophe FIORIO ^b, Michel HABIB ^{a, b}, Claude LAURENÇO ^{a, c}

^a GDR CNRS 1093, Traitement informatique de la connaissance en chimie organique, Montpellier, France

^b Laboratoire d'informatique, de robotique et de microélectronique de Montpellier, UMR CNRS 5506-UM II, 161, rue Ada, 34092 Montpellier cedex 5, France

^c Centre CNRS-Inserm de pharmacologie-endocrinologie, UPR CNRS 9023, 141, rue de la Cardonille, 34094 Montpellier cedex 5, France

E-mail: {dietz, fiorio, habib, cl}@lirmm.fr

(Received 8 October 1997, accepted after revision 3 December 1997)

Abstract – We propose to use combinatorial maps as a non-geometric computer representation of convex polyhedral and polygonal stereogenic entities. This model is simple, uniform, and general. It allows to represent any relative spatial arrangement of atoms (stereogenic entities, molecular shapes, etc.). © Académie des Sciences/Elsevier, Paris.

computer chemistry / stereochemistry / structure representation / combinatorial maps

Résumé – Représentation de la stéréochimie fondée sur la topologie. Nous proposons une représentation informatique des entités stéréogéniques polyédriques et polygonales convexes fondée sur les cartes combinatoires. C'est un modèle simple, uniforme et général. Il permet de représenter tout arrangement spatial relatif d'atomes (entités stéréogéniques, formes des molécules, etc.). © Académie des Sciences/Elsevier, Paris.

informatique chimique / stéréochimie / représentation de structures / cartes combinatoires

Version française abrégée

Le traitement informatique des structures chimiques nécessite une représentation informatique de la stéréochimie qui soit adéquate à chaque domaine d'application. Lorsque des coordonnées atomiques précises ne sont pas nécessaires, on peut utiliser une représentation *non géométrique*. Une telle représentation devrait avoir un pouvoir d'expression de l'ordre de celui des formules structurales, lesquelles décrivent *l'arrangement spatial relatif* des atomes d'une molécule [1].

La plupart des représentations non géométriques de la stéréochimie proposées pour la nomenclature ou le traitement informatique des structures chimiques (par exemple [2–5]) utilisent des modèles polyédriques ou polygonaux convexes, à partir desquels sont déduits des stéréodescripteurs représentant les entités stéréogéniques. Par exemple, pour représenter la configuration d'un carbone asymétrique, on place ses quatre atomes voisins sur les sommets d'un modèle tétraédrique et l'on en déduit un descripteur de configuration. La grande majorité des entités stéréogéniques peuvent être représentées ainsi.

Ces représentations ont été conçues pour le traitement de structures entières et ne se prêtent pas à celui des sous-structures. Un stéréodescripteur est attribuable à une entité stéréogénique si l'on dispose d'un ordre canonique sur l'ensemble des substituants impliqués. Cette précondition n'étant pas rem-

Communicated by François MATHEY.

* Correspondence and reprints.

plie dans le cas général des sous-structures chimiques, il est impossible de représenter la stéréochimie de celles-ci et d'inclure l'aspect stéréochimique dans une recherche par sous-structure dans une base de molécules ou de réactions. Un autre inconvénient de ces représentations est qu'il faut associer manuellement à chaque modèle un groupe de permutations [4] ou un ensemble de règles de nomenclature [2, 3] ad hoc afin de tenir compte des permutations d'atomes n'affectant pas l'arrangement spatial relatif de ces derniers sur les sommets du modèle. De plus, la plupart de ces méthodes ont été développées pour la chimie organique et leurs extensions en chimie inorganique ou organométallique sont limitées aux modèles les plus courants tels que l'octaèdre.

Ces difficultés sont évitées lorsqu'on représente explicitement l'arrangement spatial relatif des atomes d'une entité stéréogénique. Nous l'avons montré précédemment en proposant une solution générale pour les configurations des atomes et les conformations des liaisons [1]. Cependant, le problème des entités stéréogéniques non locales n'avait pas encore été pris en compte (par exemple les adamantanes asymétriquement substitués).

Dans cette note, nous proposons une approche plus simple et plus générale. Elle est fondée sur le modèle bien formalisé des cartes combinatoires [7, 8]. Ces dernières sont utilisées habituellement dans la conception assistée par ordinateur (CAO), où l'on compose des objets géométriques complexes (par exemple une carrosserie) à partir de simples *cellules* (sommets, arêtes, faces, volumes). Nous utilisons de telles cartes pour construire des entités stéréogéniques polyédriques ou polygonales.

Notre représentation stéréochimique se calcule à partir des coordonnées tridimensionnelles plus ou moins précises des atomes participant à une entité stéréogénique. Celles-ci peuvent être obtenues, par exemple à partir d'une formule structurale par un calcul utilisant des longueurs et angles de liaison standard. Dans une première étape la fermeture convexe [9] de cet ensemble de points est calculée. La deuxième étape consiste à représenter l'arrangement spatial relatif des sommets de ce polyèdre (polygone) convexe par une carte combinatoire (les *figures 1* et *2* montrent l'exemple d'un tétraèdre *t* et la carte correspondante *m*). Dans ce but, la convention suivante impose une orientation sur les faces du polyèdre (un polygone est considéré comme un polyèdre dégénéré sans intérieur) : l'orientation d'une face est la séquence circulaire des sommets que l'on rencontre en parcourant le bord de la face dans le sens des aiguilles d'une montre, vue de l'extérieur du polyèdre. Ensuite, on découpe le polyèdre le long de ses arêtes, éclatant ainsi chaque arête en deux *brins* (demi-arêtes orientées, symbolisées par des flèches sur la *figure 2*). Une permutation cyclique sur l'ensemble des brins représente l'orientation des faces du polyèdre. Une involution (permutation qui est son propre inverse, symbolisée par les lignes hachurées sur la *figure 2*) relie entre elles les faces orientées. Pour construire un polyèdre convexe à partir d'une carte, celle-ci est plongée dans l'espace tridimensionnel en respectant la convention d'orientation des faces. Ainsi, on obtient un polyèdre ayant le même arrangement spatial relatif des sommets que celui qui est à l'origine de cette carte.

Puisque notre représentation stéréochimique décrit explicitement l'arrangement spatial relatif des atomes d'une entité stéréogénique, elle permet d'effectuer des recherches dans des bases de molécules ou de réactions, tant par structure entière que par sous-structure. Des algorithmes pour tester l'équivalence d'entités stéréogéniques, engendrer l'image dans un miroir d'une telle entité et déterminer si deux entités sont énantiomorphes, ont été implémentés. On peut raffiner la représentation en étiquetant les brins avec de l'information sur la longueur relative des arêtes et la taille relative des angles du polyèdre. Cela permet de distinguer les polyèdres réguliers des irréguliers correspondants, ces derniers devant être employés pour représenter des entités stéréogéniques telles que celles relatives à l'atropisomérisme des biphényles ou de type allénique, par exemple.

La représentation par cartes combinatoires n'est pas limitée aux entités stéréogéniques polyédriques ou polygonales convexes. Elle s'applique a priori à tout arrangement spatial relatif d'atomes. On peut également envisager leur utilisation pour représenter la forme des molécules. Des applications potentielles de cette représentation non géométrique de la structure tridimensionnelle des molécules existent dans tous les domaines qui nécessitent un traitement informatique des structures chimiques.

1. Introduction

The computer processing of chemical structural information requires a computer representation of stereochemistry which is adequate for a given application domain. When no exact atomic coordinates are needed, a *non-geometric* representation can be used. This one should be as expressive as the structural formulae which represent the *relative spatial arrangement* of atoms in a molecule [1].

In the large majority of stereogenic entities, the involved atoms occupy the vertices of a convex polyhedron or polygon. Therefore, most of the non-geometric representations of molecular stereochemistry which have been proposed for the purpose of systematic nomenclature or the computer processing of chemical structures (e.g., [2–5]) use convex polyhedral and polygonal templates in order to derive stereodescriptors for stereogenic entities. The configuration of an asymmetric carbon atom, e.g., is represented by placing its four neighbor atoms onto the vertices of a tetrahedral template and deriving a configurational descriptor. Analogously, templates are used to represent the conformation of a bond or larger stereogenic entities which may encompass a whole molecule, e.g., an asymmetrically substituted adamantane. One approach [6] represents polyhedral clusters via linear notations derived from the corresponding canonically labeled molecular graphs which are embedded in R^3 using polyhedral templates.

These approaches have been devised for the treatment of full structures. Attribution of a stereodescriptor to a stereogenic entity (or a linear notation to a polyhedral cluster [6]) requires a canonical order of the involved atoms. However, substructures in general do not meet this precondition, so that stereochemical substructure search in a chemical database is impossible. Another inconvenience of the stereodescriptor-based approaches is that a template-specific permutation group [4] or set of nomenclature rules [2, 3] must be manually determined for each template to account for those permutations redistributing the atoms on the template vertices without affecting the relative spatial arrangement of the atoms. Moreover, most of these methods have been devised for organic chemistry and their extensions for inorganic and organometallic chemistry cover only the most common templates, e.g., the octahedron.

The above problems can be avoided if one explicitly represents the relative spatial arrangement of the atoms involved in a stereogenic entity. We have proposed a general solution for the representation of atomic configurations and bond conformations which is based on this idea [1]. However, non-local stereogenic entities have not yet been accounted for.

In this Note, we propose a simpler and more general approach to represent stereogenic entities. It is based on the mathematically well-formalized model of combinatorial maps [7, 8]. The latter are used in computer aided design (CAD) where simple *cells* (vertices, edges, faces and volumes) are assembled to form complex geometric objects, e.g., a car body. We use them to assemble convex polyhedral and polygonal stereogenic entities, but this topology-based approach is capable of representing more complex stereogenic entities and molecular shapes. It has potential applications in every domain that relies on the computer processing of chemical structural information.

2. Representing stereochemistry using combinatorial maps

Tetrahedral configurations are the most common ones encountered in organic molecules. We therefore use the enantiomorphic tetrahedra t and t' (figure 1) to illustrate how 2-dimensional maps (2-maps) allow to represent the relative spatial arrangement of the atoms forming a stereogenic entity (a mathematically rigorous presentation of combinatorial maps [7, 8] is beyond the scope of this article). We shall see how the corresponding 2-maps allow both to distinguish t from t' and to determine their enantiomorphism relationship.

The computer generation of our stereochemical representation starts from the set of (more

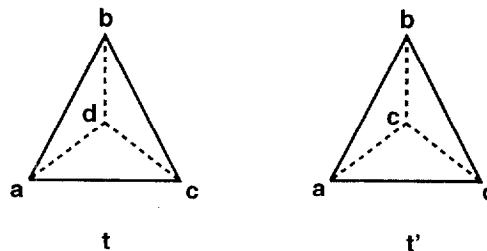


Figure 1. Enantiomorphic tetrahedra.

Figure 1. Tétraèdres énantiomorphes.

or less precise) 3-dimensional coordinates of the atoms forming a stereogenic entity. For example, the coordinates can be computed from a structural formula using standard bond lengths and angles. First, we compute the convex hull [9] of this set of points in order to obtain the corresponding polyhedron. Second, we generate the combinatorial map representing the relative spatial arrangement of the vertices of this polyhedron.

2.1. How to obtain a 2-map from a polyhedron (and vice versa)

The second step can be informally described as follows (e.g., the transition from the tetrahedron t in figure 1 to the corresponding 2-map m in figure 2).

In order to represent the relative spatial arrangement of the vertices of a polyhedron, it is necessary to impose an orientation on its faces. We adopt the convention to view each face along an axis which is perpendicular to that face and directed from the exterior to the interior of the polyhedron. Traversing the boundary of the face in clockwise direction (when viewed along the axis) yields a circular sequence of vertices defining the orientation of that face. For example, the circular sequence (a, b, c) defines the orientation of the ‘front face’ of the tetrahedron t .

We cut the polyhedron along its edges, thus obtaining the set of its oriented faces. In the case of the tetrahedron t we obtain $\{(a, b, c), (a, c, d), (a, d, b), (b, d, c)\}$. Each edge is thereby split up into two *dart*s. A dart can be thought of as a ‘directed half-edge’ whose starting point

(symbolized by a disc) and end point (symbolized by an arrowhead) coincide with polyhedron vertices. A dart is identified by a natural number and labeled with the polyhedron vertex at its start point. For example, vertex c is the label of dart 0 in m .

The oriented polyhedron faces are represented by a permutation β_1 on the set of darts. This permutation is symbolized by the circular dart sequences, where the end point of a dart x touches the start point of a dart y if and only if β_1 maps x onto y . We note a mapping $\mu: X \rightarrow Y$ as a subset of the cartesian product $X \times Y$. In m , $\beta_1 = \{(0, 1), (1, 2), (2, 0), (3, 4), (4, 5), (5, 3), (6, 7), (7, 8), (8, 6), (9, 10), (10, 11), (11, 9)\}$ induces the four circular dart sequences $(0, 1, 2), (3, 4, 5), (6, 7, 8)$, and $(9, 10, 11)$ as well as the corresponding circular label sequences $(c, a, b), (a, c, d), (b, a, d)$, and (c, b, d) , thus representing the oriented faces of t .

The oriented faces are held together by an involution β_2 on the set of darts (symbolized by thick dashed lines in figure 2). An involution is a permutation that is identical with its inverse. Two darts that are connected via β_2 correspond to a polyhedron edge. In m , $\beta_2 = \{(0, 3), (1, 6), (2, 9), (3, 0), (4, 11), (5, 7), (6, 1), (7, 5), (8, 10), (9, 2), (10, 8), (11, 4)\}$ maps, e.g., dart 0 onto dart 3 and vice versa. Thus, the dart set $\{0, 3\}$ corresponds to the polyhedron edge ac .

The inverse process of (re)constructing a polyhedron from a 2-map may be informally described as follows. The involution β_2 (thick dashed lines) might be viewed as a set of flexible connections between the oriented faces (circular dart sequences) of the 2-map. The latter are embedded into R^3 by shrinking these connec-

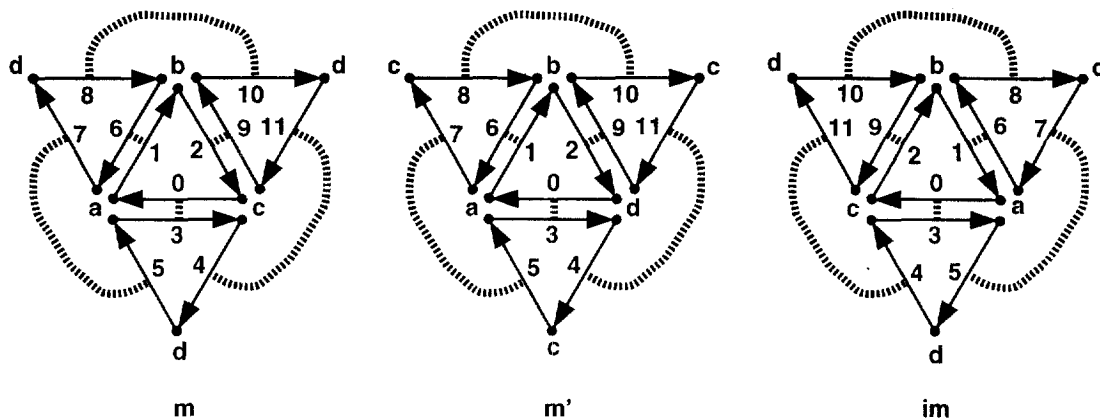


Figure 2. 2-maps representing the tetrahedra in figure 1.

Figure 2. 2-cartes représentant les tétraèdres sur la figure 1.

tions such that the starting point of each dart coincides with the end point of its image under β_2 and that the circular dart sequences are oriented clockwise, when viewed from the exterior of the resulting convex polyhedron. Thus, if we embed the oriented face $(0, 1, 2)$ of the 2-map m in the paper plane as drawn in *figure 2*, the remaining faces must be folded into the half space below the paper plane in order to give the tetrahedron t .

Definition: 2-dimensional map (2-map). A 2-dimensional map (2-map) is a quintuple $M = (D, \beta_1, \beta_2, L, \lambda)$ such that D is a finite non-empty set of darts, β_1 a permutation of D , β_2 an involution of D , L a set of dart labels, and λ a surjection from D onto L .

2.2. Comparing polyhedra via their corresponding 2-maps

Let V_P and V_Q be the vertex sets of two polyhedra P and Q and let $R_V \subseteq (V_P \cup V_Q) \times (V_P \cup V_Q)$ be an equivalence relation. By default, R_V is the identity relation $\{(v, v) \mid v \in V_P \cup V_Q\}$, i.e., the vertices of the polyhedra are distinguishable from each other. The equivalence relation may be redefined, e.g., considering vertices as equivalent if the corresponding atoms belong to the same chemical element.

Let the 2-maps M_P and M_Q represent P and Q , respectively. Because each dart of the corresponding dart sets D_P and D_Q is labeled with a polyhedron vertex from V_P and V_Q , respectively, R_V is the natural equivalence relation on the union of the sets of dart labels.

We consider P and Q as equal if M_P and M_Q are isomorphic, i.e., if there exists a bijection from D_P to D_Q which respects the adjacency of darts (as defined by the permutations β_1 and β_2) and which maps each dart from D_P onto an equivalently labeled dart of D_Q .

Definition: isomorphism of 2-maps. Let $M = (D, \beta_1, \beta_2, L, \lambda)$ and $M' = (D', \beta_1', \beta_2', L', \lambda')$ be 2-maps and let $R \subseteq (L \cup L') \times (L \cup L')$ be an equivalence relation. A bijection $\sigma: D \rightarrow D'$ is an isomorphism which maps M onto M' if and only if $\forall d \in D: \sigma(\beta_1(d)) = \beta_1'(\sigma(d)) \wedge \sigma(\beta_2(d)) = \beta_2'(\sigma(d)) \wedge (\lambda(d), \lambda'(\sigma(d))) \in R$.

Let us compare the tetrahedra t and t' (*figure 1*) which are represented by the 2-maps m and m' (*figure 2*). Accidentally, both polyhedra have the same vertex set $V = \{a, b, c, d\}$, and both 2-maps have the same dart set $D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$. If the natural dart label equivalence relation is the identity

relation $R_V = \{(v, v) \mid v \in V\}$, there exists no isomorphism mapping m onto m' , i.e., t and t' are not equal. However, if the natural dart label equivalence relation was $R_V = \{(v, v) \mid v \in V\} \cup \{(c, d), (d, c)\}$, the identity mapping on the dart set D would be an isomorphism that maps m to m' , i.e., t and t' would be equal.

2.3. How to construct the mirror image of a polyhedron

Because of the clockwise convention, inverting the orientation of the faces of a polyhedron amounts to turning the polyhedron boundary inside out and therefore leads to its mirror image.

Definition: inverse of a 2-map. Let $M = (D, \beta_1, \beta_2, L, \lambda)$ be a 2-map. The inverse of M is $M^{-1} = (D, \beta_1^{-1}, \beta_2, L, \lambda^{-1})$ with $\lambda^{-1}: D \rightarrow L$, $\lambda^{-1}(d) = \lambda(\beta_1(d))$.

Taking into account this definition and the procedure for the (re)construction of polyhedra from 2-maps, it should be clear that a 2-map M^{-1} represents the mirror image of the polyhedron represented by the 2-map M . For example, im (*figure 2*) is the inverse of m and therefore represents the mirror image t' of t (*figure 1*).

2.4. How to determine if two polyhedra are enantiomorphic

Let the 2-maps M_P and M_Q represent the polyhedra P and Q , respectively. P and Q are enantiomorphic if and only if there does not exist an isomorphism which maps M_P^{-1} onto M_P (i.e., P is chiral) and if there exists an isomorphism which maps M_P^{-1} onto M_Q (i.e., the mirror image of P and Q are equal).

It can thus be shown (using the 2-maps m , im and m' in *figure 2*) that the tetrahedra t and t' in *figure 1* are enantiomorphic.

3. Conclusions

2-Dimensional maps are a simple topology-based model providing a uniform representation for convex polyhedral and polygonal stereogenic entities (a polygon can be considered as a degenerate polyhedron which has two faces but no interior). Their scope is easily extended by including into the dart labels information about relative lengths of polyhedron edges ('vertex a is closer to vertex b than to vertex c)

and relative sizes of angles ($\angle abc$ is smaller than $\angle bcd$, but it is approximately the size of dab). This is necessary to distinguish regular from irregular polyhedra (e.g., stereogenic entities of the allene type or arising from atropisomerism in biphenyls may be represented using an irregular tetrahedron).

The 2-map representation of stereochemistry describes explicitly the relative spatial arrangement of the atoms of a stereogenic entity and therefore allows to perform full structure search as well as substructure search in chemical databases. Algorithms have been implemented which determine the equivalence of stereogenic entities, construct the mirror image of such an entity, and determine if two entities are enantiomorphic. Recognition of stereochemical relationships (*cis/trans*, *syn-/antiperiplanar*, *+/-synclinal*, *+/-anticlinal*) has also been implemented.

References

- [1] Dietz A., J. Chem. Inf. Comput. Sci. 35 (1995) 787–802.
- [2] Cahn R.S., Ingold C.K., Prelog V., Angew. Chem. Int. Ed. Engl. 5 (1966) 385–415.
- [3] (a) Petrarca A.E., Rush J.E., J. Chem. Doc. 9 (1969) 32–37. This method has subsequently been used by other authors, notably by Wipke, Corey and Jorgensen.
(b) Wipke W.T., Dyott T.M., J. Am. Chem. Soc. 96 (1974) 4825–4834.
(c) Corey E.J., Howe W.J., Pensak D.A., J. Am. Chem. Soc. 96 (1974) 7724.
(d) Peishoff C.E., Jorgensen W.L., J. Org. Chem. 50 (1985) 3174–3184.
- [4] Ugi I., Dugundji J., Kopp R., Marquarding D., Perspectives in Theoretical Stereochemistry, Lecture Notes in Chemistry 36, Springer-Verlag, Berlin, 1984.
- [5] Dubois J.-E., Alliot M.-J., Panaye A., C. R. Acad. Sci. Paris 273 (1971) 224–227.
- [6] Herndon W.C., Leonard J.E., Inorg. Chem. 22 (1983) 554–557.
- [7] Lienhardt P., Comput. Aided Design 23 (1991) 59–81.
- [8] Cori R., Astérisque 27 (1975).
- [9] Preparata F.P., Shamos M.I., Computational Geometry: An Introduction, Springer-Verlag, New York, 1985.
- [10] Walba D.M., Tetrahedron 41 (1985) 3161–3212.

The combinatorial map model allows to easily extend the scope of our representation to cover the few stereogenic entities not amenable to 2-maps: the relative spatial arrangement of any set of points in R^3 can be represented by a 3-map which is obtained by 'sewing' 2-maps. Thus, 3-maps should also lend themselves to the representation of molecular shapes. Molecular shapes corresponding to non-orientable surfaces (e.g., a molecular Möbius strip [10]) can be represented using generalized n -dimensional maps [7]. Further investigation in these areas is underway.

Acknowledgement

The work of one author (A.D.) has been financially supported by Grant N° 96/10443 of the French Ministry of Foreign Affairs. This support is gratefully acknowledged.