The new molecular surface descriptors

A.M. Shestov, **A.V. Koptsova, M.I. Kumskov**

Department of Computational Mathematics, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, 119992 Moscow, Russian Federation.
E-mail: qsar_msu@mail.ru, shestov.msu@gmail.com

ABSTRACT

A new method for extracting geometric features (we call them singular points) of triangulated molecular surface is proposed. For this purpose we developed an algorithm of molecular surface segmentation, which combines several existing techniques. New 3d structural descriptors (we will call them the molecular surface descriptors) are constructed on the base of singular points, which take into account a geometry of molecular surface and local physico-chemical properties of the compound. These descriptors are used for the solution of the QSAR problem. The algorithm is developed and its complexity and needed memory are estimated. The project is implemented on the C++ language and tested on the 2 training sets – the glycosides and the toxic compounds.

Keywords: *QSAR, Descriptors, Molecular Surface, Segmentation.*

1. INTRODUCTION

1.1 QSAR problem definition

The QSAR problem is as follows: to determine the relationship between the structure of chemical compounds and the properties [8]. The solution of this problem can be presented in two stages: the choice of the molecular graphs description and the discriminant function construction. The second stage can be solved by general methods of the pattern recognition [15]. The difference between the QSAR problem and other problems of pattern recognition is the first stage of the solution (the molecular graphs description). The molecular graphs description is the problem which this paper is devoted to.

Let us consider the basic definitions.

The molecular graph is a singly labeled graph whose vertices are interpreted as atoms of the molecule, and the edges – as covalent bonds between pairs of atoms. Vertices and edges may have additional attributes such as coordinates for the vertices, or information about whether a connection is ring. Labels of vertices are the names of the atoms in the periodic table, labels of the edges are the types of bonds between atoms (single, double, triple, aromatic). Each vertex of the graph corresponds to the three real numbers – the location of the atom in some coordinate system.

The descriptor is any property which numerical value can be computed for any molecular graph.

Consider for each atom A_i of a molecule G a ball $B_o(V_i, r_i)$, which radius r_i is a **Van der Waals radius** [16] of the atom V_i and the center V_i coincides with the center of A_i. **The Van der Waals (VDW) surface *V*(*G*)** is the border of the union of these balls (see fig. 1): $V(G) = \partial \bigcup_{i=1}^{n} B_o(V_i, r_i)$

Consider a probe sphere ***S*** of radius ***r***. This sphere is "rolling" over the **Van der Waals surface *V*(*G*). The molecular surface** (or **Connolly surface**) ***M***$\boldsymbol{M_r(G)}$ [2] of radius **r** of a molecule ***G*** is the surface that is accessible to a probe sphere of the radius ***r*** rolling over the **Van der Waals surface** (see fig. 1):

$$M_r(G) = \partial\{x \in R^3 : \exists y \notin \bigcup_{i=1}^{n} B(P_i, r_i), \rho(x,y) \le r\}$$

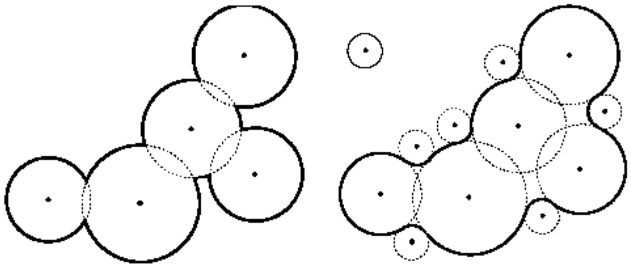


Fig. 1 VDW (left) and molecular (right) surfaces

1.2 Definition of the problem of the molecular surface descriptors construction

Decide that we are given a chemical compound ***M*** and its triangulated molecular surface ***G***. The problem is as follows: to develop a way of describing a molecular surface ***G*** with a set of singular points $\boldsymbol{\Sigma}=\{\boldsymbol{\xi}_i\}$, $i = 1,\ldots,n$, which fulfill the following conditions:

1. Every singular point $\boldsymbol{\xi}_i$ must refer to some geometric feature $\boldsymbol{g_j}$ of the surface ***G***, e.g. *the description must be informative.*
2. Every geometric feature $\boldsymbol{g_j}$ must refer to one and only one singular point $\boldsymbol{\xi}_i$, e.g. *the description must be total and nonredundant.*

Then, decide we are given a training set $\boldsymbol{L}=\{\boldsymbol{M_i}\}$, *i=1,…,m.*, where each compound $\boldsymbol{M_i}$ have a triangulated molecular surface $\boldsymbol{G_i}$**.** When a set of singular points $\boldsymbol{\Sigma}_i$ is constructed for each molecule $\boldsymbol{M_i}$ of the training set, the alphabets of descriptors (we will call them the molecular surface descriptors further) are constructed in the following way:

1. Several local physical or chemical properties are calculated in each singular point $\boldsymbol{\xi}_j$ for each molecule $\boldsymbol{M_i}$ (e.g. electrostatic potential, lipophilicity, hydrophoby, molecular fields [5]).
2. Each singular point is classified according to the type of singularity, which it refers to, and the value of calculated

property: $\boldsymbol{l_j} = T(\boldsymbol{\xi}_j)$. This is the way the alphabet of descriptors A^1 is calculated.

3. Each pair of singular points (ξ_i, ξ_j) is classified according to types $T(\boldsymbol{\xi}_j)$, $T(\boldsymbol{\xi}_i)$ and the type of distance between them $T(||\boldsymbol{\xi}_j, \boldsymbol{\xi}_i||)$: $\boldsymbol{l_{ij}} = T(\boldsymbol{\xi}_j, \boldsymbol{\xi}_i)$. This is the way the alphabet of descriptors A^2 is calculated.
4. Each triple of singular points (ξ_i, ξ_j, ξ_l) is classified according to types $T(\boldsymbol{\xi}_j)$, $T(\boldsymbol{\xi}_i)$, $T(\boldsymbol{\xi}_l)$ and the type of triple of distances between them $T(||\xi_i, \xi_j||, ||\xi_i, \xi_l||, ||\xi_j, \xi_l||)$: $\boldsymbol{l_{ijl}} = T(\xi_i, \xi_j, \xi_l)$. This is the way the alphabet of descriptors A^3 is calculated.

The constructed alphabets are used for the predicting activity of the new compounds.

1.3 The topicality of the problem of the molecular surface descriptors construction

The activity of a molecule is encoded in its shape. Of all geometric properties of a molecule, its surface play an essential role as it delineates the region covered by the protein and therefore defines its region of interactions. Characterization of the compound's molecular surface therefore plays an essential role for analyzing and predicting biomolecular complexes, as well as for modeling the energetic of formation of such complexes. As the surface of a molecule also defines its interface with the solvent it bathes in, its form is crucial for understanding the salvation [9].

So, a set of singular points $\boldsymbol{\Sigma}=\{\boldsymbol{\xi}_i\}$, classified according to the neighborhood shape and physico-chemical properties, is informative from the chemical point of view.

One of the ligand-receptor interaction models is a "triangle of activity" [8]. *So, a set of pairs and triples of singular points, classified according to types of singular points and type of distances between them, is also informative from the chemical point of view.*

Also the developed method of molecular surface segmentation can be used for docking [6].

2. THE DESCRIPTION OF THE ALGORITHM

2.1 The idea of the singular points extraction

The idea is to divide a molecular surface G into the set of segments $\{S_i\}$, $i = 1\ldots n_s$, in such a way, that each segment S_i is a region of a specific shape – convex, concave or saddle. Then we take centers of constructed segments as singular points. Because each region of a specific shape S_i contains one and only one geometric feature, the obtained set of singular points will fulfill the conditions formulated in chapter 3. This idea is new approach for extracting geometric features.

So, the key question is how to segment a molecular surface.

2.2 The general approach to the segmentation of triangulated surfaces

Let us consider general approach for segmenting triangulated surface.

Decide we are given a triangulated surface $\boldsymbol{G}$, which consists of a set of vertices $\boldsymbol{V}=\{\boldsymbol{v_i}\}$, $i=1\ldots n_v$ and a set of triangles $\boldsymbol{T}=\{\boldsymbol{t_i}\}$, $i=1\ldots n_t$. The most common approach for segmenting a triangulated surface G can be presented in two stages:

1. The choice of scalar function $F{:}G\rightarrow\boldsymbol{R}$ and its calculation for each vertex $\boldsymbol{v_i}$.
2. Segmentation of the surface according to the value of F.

The methods of segmentation can be generally divided into 2 types – *feature-based segmentation* and *part-based segmentation*. For our purpose we need the *feature-based segmentation.*

Let us consider the first stage. The most frequently used functions are mean and Gaussian curvatures and some their combinations. We can't compute these curvatures directly, because only triangulation of the molecular surface is given. So we need to estimate it. A review of different methods for the principal directions estimation can be found in [4]. The segmentation method based on the curvature can be found in work [14] . Also there were developed functions for segmenting the molecular surface precisely – Connolly function [1] and atomic density function [10].

Let's consider the second stage. There are 2 general approaches – *extracting regions* and *extracting borders.* When we extract regions, we unite elements of surface (vertices or triangles) with similar values of F into same regions. The examples can be found in [7][14] . When we extract borders, we treat elements of surface (vertices or triangles) with high values of F as borders between regions. The examples can be found in [10][1] .

In this work we use Haussian and mean curvature as F and after that extracted regions. The choose of curvature is explained by two facts –

1. Due to the method of construction of the molecular surface curvature perfectly suits for its segmentation. This is explained in the next chapter.
2. Unlike the Connolly function and the atomic density function curvature is easy calculated.

2.3 The idea of the triangulated molecular surface segmentation

The molecular surface is composed of fragments of spheres, toruses and reentrant surfaces of spheres. A part of sphere is obtained when a probe sphere touches only one sphere of VDW surface, a part of torus – when a probe touches 2 spheres, a part of reentrant surface of sphere – when a probe touches 3 spheres [12].

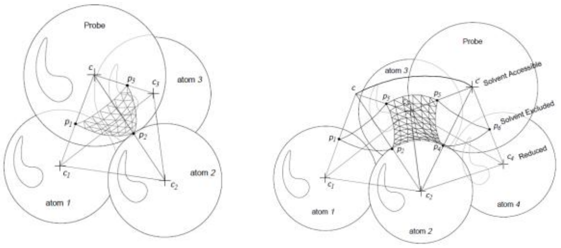


Fig. 2 Torus and reentrant surface of sphere

So, parts of different spheres of WDV surface are separated by parts of toruses, parts of reentrant surfaces of spheres are separated by parts of toruses too.

These 3 different types of molecular surface compounds (spheres, toruses and reentrant surfaces of spheres) can be distinguished by the sign of their Gaussian and mean curvatures. Denote Gaussian curvature as $\boldsymbol{K}$ and mean curvature as $\boldsymbol{H}$. Then a sphere has $K>0$, $H>0$, torus - $K<0$, reentrant surface of sphere - $K>0$, $H<0$.

So, the idea of algorithm is:

1. We assign each vertex to one of 3 types (*convex, concave, saddle*) according to its Haussian and mean curvature.

2. Extract regions of spheres (*convex regions*), regions of reentrant surface of spheres (*concave regions*), as they are separated by regions which have negative Gaussian curvature.

2. Regions which have negative Gaussian curvature (*saddle regions*) in most cases form a small number of connectivity components. So they can be segmented by the algorithm of clustering *k-means* [11].

The strong point of the proposed idea is the simplicity of calculations – we need only a sign of curvature, not its value. Because of this the result of this algorithm must suffer less from irregularities in triangulation of a surface.

2.4 The implementation of the algorithm of segmentation

2.4.1 Molecular surface construction

For the molecular surface constructing we used the program Accelrys Discovery Studio 2.5 [17]. It differs from the other programs for the molecular surface construction with 2 facts –

1. The quality of triangulation is better than in other programs.
2. The triangulated surface can be saved in *wrl* format, which provides not only vertices and triangles, but also a normal to surface in each vertex. So, we have no need to estimate surface normals. And estimation of normals is not a simple question, many articles are devoted to this problem.

2.4.2 The curvature calculation

We chose 2 algorithms for curvature estimation – estimation by a parabolloid [4] and estimation by a cubic-order surface [4]. The first is used for the Gaussian curvature calculation and the second – for the mean curvature calculation.

The essence of these methods is the following: we try to estimate the neighborhood surface of the vertex of the triangulation by a parabolloid or by a cubic-order surface. We obtain a overdetermined system of linear equations, which we solve by the least-squares method. The result is the analytical form of the parabolloid or the cubic-order surface. Then we calculate principal curvatures, Haussian(K) and mean(H) curvatures of the obtained surface. Then we assign each vertex to one of three types: convex (K>0, H>0), saddle (K<0), concave (K>0, H<0).

There occur errors because of the irregularities in triangulation, such that for the vertex v_i of type k_1 all adjacent vertices v_j belong to another type k_2.We use the following solution: we assign vertex v_i to the type k_2.

2 another types of errors are considered in the next chapter.

After this stage of algorithm all vertices of triangulated surface belong to one of 3 types: concave, convex or saddle.

2.4.3 Convex and concave regions extraction

There are 2 variants of segmentation – to take vertices as elements of segments or to take triangles. If we take triangles we must develop a way to classify them as we did it for vertices. We do it in the following way: triangle t is assigned to a convex (concave) type if and only if all its vertices v_{t1}, v_{t2}, v_{t3} have the convex (the concave) type. Otherwise t is assigned to a saddle type.

This procedure helps to fix the following error of curvature calculation: some convex or concave regions which are meant to be disconnected become connected. If we take triangles as elements of segments different concave (convex) regions will be disconnected.

2.4.4 Saddle regions segmentation

Regions with K<0 (*saddle regions*) in most cases form a small number of connectivity components. Each connectivity component is segmented distinguishly. We tried 2 algorithms for segmentation – *k-means* [11] and *mountain clustering* [13]. In this stage we take vertices as elements of segments. For each algorithm we need to define distances between vertices. We take the shortest geodesic distance as distance between vertices. Because of it each obtained cluster has only one component of the connectivity and can be treated as a segment. The shortest distances are calculated by the Dijkstra algorithm with a binary heap as a priority queue [3]. *k-means* provides better segmentation, so we chose it instead of the *mountain clustering*.

After this step the surface is totally segmented.

2.4.5 Singular points extraction

In each segment we take as a singular point the vertex which is the closest to the geometric center of a segment:

$$\xi_i = arg\ min_{v_j \in S_i} \left\| v_j - \frac{1}{n_i} \sum_{x_m \in S_i} v_m \right\|$$

We can see segmented molecular surface with singular points on the figure.

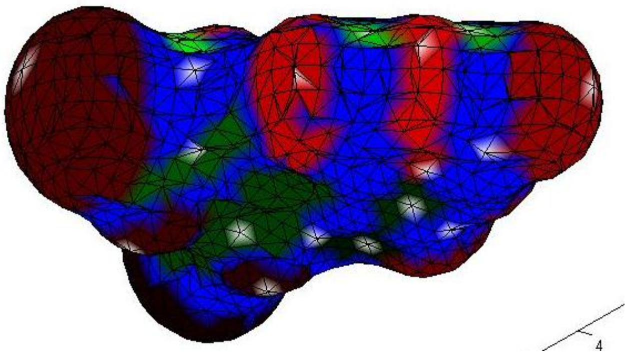
3. THE COMPLEXITY OF THE ALGORITHM OF SINGULAR POINTS EXTRACTION

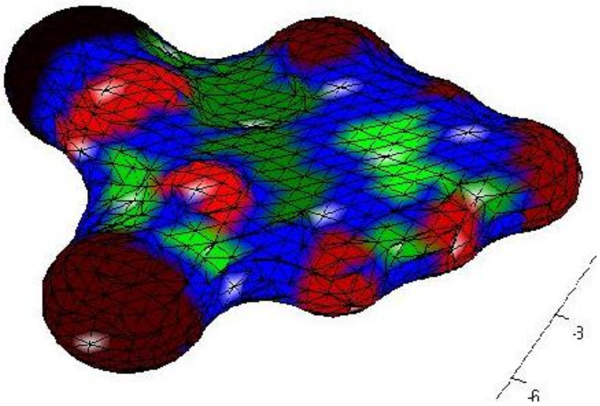
The complexity is - $\boldsymbol{\Theta(N_t^{2} ln\ N_t)}$.

The memory which we need - $\boldsymbol{\Theta(N_t^{2})}$, where N_t is a number of triangles in the triangulation of the molecular surface.

4. THE RESULTS

On the figures you can see segmented molecular surfaces with singular points.





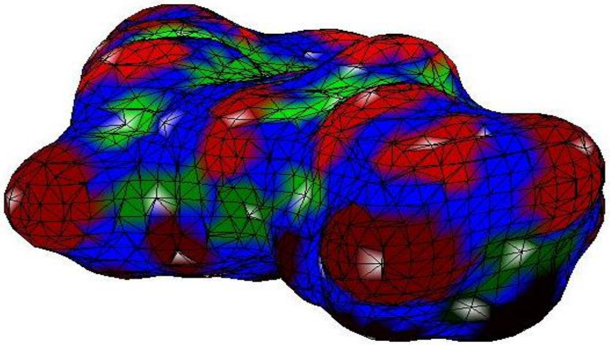


Fig. 3 Segmented molecular surfaces with singular points

On the base of singular point we constructed the alphabet of descriptors A^2 (see the definition). The constructed descriptors were tested on the 2 training sets – the glycosides and the toxic compounds. For each set we constructed 8 molecular-descriptor matrices [8]. We took a percentage of correctly classified compounds as the prediction quality. The results are the following:

№	The prediction quality	Ejections
1	0.7962	2
2	0.8454	0
3	0.8113	4
4	0.8504	3
5	0.8504	3
6	0.7909	0
7	0.8055	2
8	0.8545	0

Table 1 The glycosides

№	The prediction quality
1	0.49
2	0.48
3	0.44
4	0.51
5	0.52
6	0.45
7	0.49
8	0.51

Table 2 The toxic compounds

5. THE CONCLUSION

The algorithm of describing a molecule with a set of singular points is proposed. The complexity of the algorithm and the needed memory are estimated. 3 alphabets of descriptors are constructed on the base of singular points. Experiments show, that these descriptors a suitable for predicting the activity of new compounds.

The further direction of the work is to build a fuzzy classification of singular points. Then the prediction quality will continuously depend on the initial parameters of the algorithm.

6. ACKNOLEDGMENTS

This work is supported by the Russian Foundation for Basic Research, project no. 10–07–00694.

7. REFERENCES

1. F. Cazals, F. Chazal, T. Lewiner: "Molecular Shape Analysis based upon the Morse-Smale Complex and the Connolly Function”. *Annual Symposium on Computational Geometry, Proceedings of the nineteenth annual symposium on Computational geometry, San Diego, California, USA, SESSION: Topology* (2003), 351 – 360.
2. M. Connolly: "Analytical molecular surface calculation", *J. Appl. Crystallogr*., i.16 (1983), 548–558.
3. T. Corman, C. Leiserson: "Introduction to Algorithms"
4. J. Goldfeather, V. Interrante: "A Novel Cubic-Order Algorithm for Approximating Principal Direction Vectors", *ACM Transactions on Graphics (TOG)*, v. 23, i. 1 (2004) .
5. T. Halgren: "Merck molecular force field", *J.Comp Chem*, i. 17 (1996), 490-641.
6. H. Holtje, W. Sippl, D. Rognan, G. Folkers: "Molecular modeling. Basic principles and applications".
7. G. Lavoue, F. Dupont, A. Baskurt: "A new CAD mesh segmentation method, based on curvature tensor analysis", *Computer-Aided Design*, i. 37 (2005), 975–987.
8. G. Makeev, M. Kumskov , I. Svitan’ko , I.Zyryanov: "Recognition of Spatial Molecular Shapes of Biologically Active Substances for Classification of Their Properties", *Pattern Recognition and Image Analysis*, v.6, i.4 (1996), p.795-808.
9. V. Natarajan, P. Koehl, Y. Wang, B. Hamann: "Visual Analysis of Biomolecular Surfaces", *Mathematics and Visualization*, i.5 (2008), 237-255.
10. V. Natarajan,Y. Wang, P.-T. Bremer,V. Pascucci, B. Hamann "Segmenting molecular surfaces", *Computer Aided Geometric Design*, i. 23 (2006), 495–509
11. A. Perevoznikov, A. Shestov, E. Permyakov, M.Kumskov: "A Way to Increase the Prediction Quality for the Large Set of Molecular Graphs by Using the *k*–NN Classifier", *Pattern Recognition and Image Analysis*, v. 21, i.2 (2011).
12. M. Sanner, A.Olson, J. Spehner.: "Reduced Surface: an Efficient Way to Compute Molecular Surfaces" *Biopolymers*, v. 38, i.3 (1996), 305-320.
13. S.Shtovba: "Introduction into fuzzy logic", 12.2.1.
14. T. Srinark, C. Kambhamettu: "A novel method for 3D surface mesh segmentation", *Proceedings of the 6th IASTED International Conference on Computers, Graphics, and Imaging* (2003), 212-217.
15. G. Tu, R. Gonsales: "Principles of pattern recognition", ed. "Mir" (1978).
16. U. Zefirov, P. Zorkiy : "Van der Waals radii of atoms In Crystal Chemistry and Structural Chemistry (History)". *Problems of Crystal Chemistry*, ed. "Science" (1992), 6-24.
17. http://accelrys.com/products/discovery-studio/visualization-download.php