JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 11, 2017
Mary Ann Liebert, Inc.
Pp. 1134–1137
DOI: 10.1089/cmb.2017.0061

Cavity Versus Ligand Shape Descriptors: Application to Urokinase Binding Pockets

NATACHA CERISIER,[1] LESLIE REGAD,[1] DHOHA TRIKI,[1]
ANNE-CLAUDE CAMPROUX,[1] and MICHEL PETITJEAN[1,2]

ABSTRACT

We analyzed 78 binding pockets of the human urokinase plasminogen activator (uPA) catalytic domain extracted from a data set of crystallized uPA–ligand complexes. These binding pockets were computed with an original geometric method that does *NOT* involve any arbitrary parameter, such as cutoff distances, angles, and so on. We measured the deviation from convexity of each pocket shape with the pocket convexity index (PCI). We defined a new pocket descriptor called distributional sphericity coefficient (DISC), which indicates to which extent the protein atoms of a given pocket lie on the surface of a sphere. The DISC values were computed with the freeware *PCI*. The pocket descriptors and their high correspondences with ligand descriptors are crucial for polypharmacology prediction. We found that the protein heavy atoms lining the urokinases binding pockets are either located on the surface of their convex hull or lie close to this surface. We also found that the radii of the urokinases binding pockets and the radii of their ligands are highly correlated ($r = 0.9$).

Keywords: algorithms, protein binding pockets, statistics, urokinases.

1. INTRODUCTION

Several thousands of molecular descriptors are known (Todeschini and Consonni, 2008). Although they are suitable for protein ligands, most of them are meaningless for protein pockets. This is particularly true for geometric descriptors, because the ligand shape is commonly associated with some envelope separating the ligand to its exterior, whereas the shape of a protein pocket is rather associated with the boundary of a cavity internal to the protein. Thus, we used our own pocket descriptors (Section 3). Then, the definition of a cavity inside a protein is highly polemical: see the many algorithms cited by Pérot et al. (2010)

[1]MTi, INSERM UMR-S 973, Université Paris Diderot, Paris, France.
[2]Epôle de Génoinformatique, Institut Jacques Monod, CNRS, UMR7592, Université Paris Diderot, Paris, France.

and by Benkaidali et al. (2014). The main problem encountered in pocket calculation algorithms, and more generally in modeling algorithms, is the existence of arbitrary parameters having a crucial effect on the results. It is why we built our own pocket calculation algorithm, which is parameter free (Section 2.2). We applied this calculation to a set of human urokinase plasminogen activator (uPA) catalytic domains. It is an attractive therapeutic target in cancer because it plays an essential role in the process of tumor cell migration and metastasis (Andreasen et al., 2000). Furthermore, the uPA receptor system is known as a strategic therapeutic target (Degryse, 2013).

2. METHODS

2.1. Data preparation

From the Protein Data Bank (PDB; Berman et al., 2000), we extracted a set of 97 crystallized uPA catalytic domain–ligand complexes. To remove nonspecific and nonbiological ligands, we removed the crystallization additives and salts. We also removed all hydrogen atoms to get a homogeneous set. In the case of polymers with multiple ligands (e.g., 2VNT PDB code), we duplicated the file in the ones where there are ligands, to launch an automated treatment for pocket calculations. The final working set contained 71 human urokinase catalytic domains, from where 78 pockets are extracted, each pocket containing one ligand.

2.2. Pocket calculation algorithm

We define a protein pocket as protein atoms extracted using the two following steps: (1) for each ligand atom we retain its closest neighboring atom in the protein and (2) in the case of multiple copies of a protein atom, we retain only one to get a nonredundant set. The drawback of this algorithm is that it cannot apply if there is no ligand. However, the strong advantage of our algorithm is its simplicity, particularly the fact that it does not require any parameter.

3. RESULTS AND DISCUSSION

We calculated the pocket descriptors of Table 1 with the *PCI* freeware. The 78 PCI values ranging from 0 to 0.04 indicate highly convex pockets. The pocket sphericity index values range from 0.14 to 0.50: the largest inscribed sphere (radius R_i) is smaller than the radius of the smallest circumscribed sphere (radius R_h): the pockets are a bit flat. It is stressed that a pocket could be nearly flat while its bounding atoms indeed lie on the surface on a sphere. The distributional sphericity measured by the distributional sphericity coefficient (DISC) parameter shows to which extent the pocket protein atoms lie on the surface of a sphere (Appendix). The DISC values range from 0.01 to 0.27. Thus, the pocket shapes are moderately fitted by spheres. To study the correspondence between pocket and ligand's shapes, we looked at the correlation between the pocket radii R (Appendix) and the radii R_{hL} of the smallest sphere enclosing the ligand. We found a high correlation coefficient $r(R, R_{hL}) = 0.89$. The correlation coefficient $r(R_h, R_{hL}) = 0.91$ is also high. These results show that our method is suitable to estimate pockets guided by the ligand. This

Table 1. The Main Pocket Descriptors Computed with *PCI*

N	Number of pocket atoms
R^h	Radius of the convex hull of the N atoms (Petitjean, 1992)
R^i	Radius of the largest sphere inscribed in the convex hull
PSI	Pocket sphericity index[a]: R_i/R_h; takes values in [0;1]
PCI	Pocket convexity index[a,b]; takes values in [0;1]
R	Pocket radius (see Theorem 4.1 in Appendix)
DISC	Distributional sphericity coefficient (Appendix)

[a]These indices were first mentioned by Borrel et al. (2015).
[b]Ratio of the squared quadratic mean distance of the N atoms to their hull, to R_i^2.

estimation is a crucial step to predict interaction partners and off targets, and to address polypharmacology (Abi Hussein et al., 2017). Our combined use of descriptors with a free parameter pocket estimation permitted us to evaluate the adaptation of the pocket to the size of the ligand.

4. APPENDIX: CALCULATION OF THE BEST FITTING SPHERE

Consider n points $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^d$, $d \geq 1$. The best fitting sphere is defined so that its center c minimizes the variance V of the population of the squared distances of the n points to c. The radius R of the best fitting sphere is the square root of the mean of these n squared distances. When the minimized variance is null, all the n points lie on the boundary of a sphere of radius R centered on c.

The calculation of R and c are done according to Theorem 4.1. For convenience, this theorem is presented for random vectors. The case of n points in $\mathbb{R}^d$ is retrieved through a finite discrete random vector. In what follows, the quote denotes the transposition operator and the symbol E denotes the expectation operator.

Let X be a random vector taking values in $\mathbb{R}^d$. The random variable expressing the squared length of $X-c$ is $Z=(X-c)'(X-c)$. The variance of Z is $V=E(Z-EZ)^2$, assumed to exist. The variance matrix of X is $\mathbf{K}=EYY'$, with $Y=X-EX$. $\mathbf{K}$ is assumed to be of full rank. V_0 is the variance of $Y'Y$. We set $\gamma=EYY'Y$ and $\tau=c-EX$.

Theorem 4.1. *The center of the best fitting sphere is* $c=EX+\mathbf{K}^{-1}\gamma/2$ *and its squared radius* $R^2=EY'Y+\tau'\tau$. *The minimized variance is* $V=V_0-\gamma'\mathbf{K}^{-1}\gamma$.

Proof. $Z=(Y-\tau)'(Y-\tau)$ and $Z-EZ=Y'Y-EY'Y+\tau'\tau$. The variance of $Z-EZ$ is $V=V_0-4\tau'\gamma+4\tau'\mathbf{K}\tau$. The gradient of this variance with respect to τ is $\nabla V=-4\gamma+8\mathbf{K}\tau$. Thus we deduce that the optimal value of τ is $\mathbf{K}^{-1}\gamma/2$. The minimal variance and the optimal radius are deduced from the latter. ■

The support of X lies on the boundary of a sphere if and only if $V=0$. When $V>0$ we need to evaluate how the distribution of X deviates from this ideal case. Having $c=EX+\mathbf{K}^{-1}\gamma/2$, we could look at the normalized variance $\Delta=V/R^4$. Unfortunately, Δ can be arbitrarily large. Thus we define the quantity DISC, which takes values in [0;1]:

Definition 4.1. $\mathrm{DISC}=\Delta/(1+\Delta)$ *is the distributional sphericity coefficient.* $\mathrm{DISC}=V/(V+R^4)$.

DISC is null if and only if the support of X lies on the boundary of a sphere. The larger DISC is, the more the distribution of X deviates from this ideal situation. DISC is insensitive to isometries and scaling.

When $d=1$, the value $\mathrm{DISC}=0$ is reached if and only if the random variable X follows a Bernouilli distribution. Thus, when $d=1$, DISC may also be viewed as a bimodality coefficient.

Still when $d=1$, we retrieve in corollary 4 an interesting inequality, which goes back to Pearson (1929). Let $\mathcal{S}$ be the skewness of X, that is, its reduced centered third order moment, and $\mathcal{K}$ its kurtosis, that is, its reduced centered fourth order moment, assumed to be existing.

Corollary 4.1. $\mathcal{K} \geq \mathcal{S}^2+1$.

Proof. Set $d=1$ in Theorem 4.1. $V=\sigma^4(\mathcal{K}-1-\mathcal{S}^2)$, σ being the standard deviation of X. Write that $V \geq 0$. ■

This inequality was also mentioned by Petitjean (2013), as a consequence of a result about geometric docking (Petitjean, 2004). Theorem 4.1 can also be viewed as a consequence of this result of Petitjean (2004).

AVAILABILITY OF *PCI*

Free binaries and documentation of *PCI* are available through a software repository located at http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abi Hussein, H., Geneix, C., Petitjean, M., et al. 2017. Global vision of druggability issues, applications and perspectives. *Drug Discov. Today*. 22, 404–415.

Andreasen, P.A., Egelund, R., and Petersen, H.H. 2000. The plasminogen activation system in tumor growth, invasion, and metastasis. *Cell. Mol. Life Sci.* 57, 25–40.

Benkaidali, L., André, F., Maouche, et al. 2014. Computing cavities, channels, pores and pockets in proteins from non spherical ligands models. *Bioinformatics*. 30, 792–800.

Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.

Borrel, A., Regad, L., Xhaard, H., et al. 2015. PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties. *J. Chem. Inf. Model.* 55, 882–895.

Degryse, B. 2013. Editorial: The urokinase receptor system as strategic therapeutic target: Challenges for the 21st century. *Curr. Pharm. Des.* 17, 1872–1873.

Pearson, K. 1929. Editorial note to Inequalities for moments of frequency functions and for various statistical constants, by J.Shohat. *Biometrika*. 21, 361–375.

Pérot, S., Sperandio, O., Miteva, M.A., et al. 2010. Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discov. Today*. 15, 656–667.

Petitjean, M. 1992. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* 32, 331–337.

Petitjean, M. 2004. From shape similarity to shape complementarity: Toward a docking theory. *J. Math. Chem.* 35, 147–158.

Petitjean, M. 2013. The chiral index: Applications to multivariate distributions and to 3D molecular graphs, 11–16. *In* Zadnik Stirn, L., Žerovnik, J., Povh, J., Drobne, S., and Lisec, A., eds. *Proceedings of SOR'13, the 12th International Symposium on Operational Research in Slovenia*. Slovenian Society INFORMATIKA (SDI), Section for Operations Research (SOR), Ljubljana, Slovenia.

Todeschini, R., and Consonni, V. 2008. *Handbook of Molecular Descriptors*. Wiley, New York.

Address correspondence to:
Dr. Michel Petitjean
MTi, INSERM UMR-S 973
Université Paris Diderot
35 rue Hélène Brion
75205 Paris Cedex 13
France

E-mail: michel.petitjean@univ-paris-diderot.fr;
petitjean.chiral@gmail.com