Assessing the geometric diversity of cytochrome P450 ligand conformers by hierarchical clustering with a stop criterion.

Jamel Eddine Meslamani, François André, Michel Petitjean

ARTICLES

Assessing the Geometric Diversity of Cytochrome P450 Ligand Conformers by Hierarchical Clustering with a Stop Criterion

Jamel Eddine Meslamani,[†,‡] François André,[†] and Michel Petitjean*[,†]

CEA/DSV/iBiTec-S/SB[2]SM (CNRS URA 2096), 91191 Gif-sur-Yvette, France, and UFR de Chimie, Université Louis Pasteur, 67008 Strasbourg Cedex, France

Received August 7, 2008

An algorithm is presented, which exhibits a computed number of rigid conformers of an input small molecule, covering the geometric diversity in the conformational space, with minimal structural redundancy. The algorithm calls a conformer generator, then performs an agglomerative hierarchical clustering with the modified clustering gain as the stop criterion. The number of classes is computed without an arbitrary parameter. A representative conformer is selected in each class, and nonrepresentative conformers are discarded. For illustration, the algorithm has been applied on a database containing 70 ligands of the cytochrome CYP 3A4, showing that the structural flexibility of each ligand is indeed handled via a small number of its representative conformers. The method is valid for all small molecules.

1. INTRODUCTION

The cytochrome CYP 3A4 is a member of the P450 hemoprotein superfamily lying in the human liver. It has been estimated that it contributes to the metabolism of roughly 50% of the drugs on the market.[1] The CYP 3A4 ligands are structurally diverse compounds of relatively high molecular weight,[2] most of them offering a wide conformational diversity due to the flexibility around their rotatable bonds. The active site of the CYP 3A4 is buried inside the enzyme, so that it may be assumed that flexible ligands are submitted to conformational changes at low energetic cost along the trajectories inside the enzyme channels, and the bioactive conformations are themselves not ensured to correspond to minimal energy conformers. In the framework of in-silico studies of several human isoforms of the P450s family,[3,4] we need to consider each ligand of the CYP 3A4 as being the union of several rigid conformers for virtual screening purposes. Although recent QSAR studies of several P450s substrates consider only one conformer,[5] recent virtual screening studies involve a number of conformers to be fixed by the user, such as 50 or 100 conformers.[6,7] Neither a single rigid conformer nor an arbitrary fixed number of conformers are satisfactory for our modeling purposes, which rely on sophisticated geometrical shape analysis of the ligands along the trajectories in the CYP 3A4 channels: the number of conformers to consider should depend on the flexibility of the ligand, rather than being a fixed number. It is why we are looking to compute an optimal number of conformers of each ligand with minimal geometric redundancy, but representative of its geometric diversity within an energetically acceptable deviation from the computed minimal energy conformer. It is expected that this approach will permit to create an extended potential ligands database in order to perform a realistic high throughput virtual screening.

2. METHODS

The selection of a set of conformers of a molecule is completed in three steps: (a) generating a large set of n conformers covering the structural diversity of geometries of the molecule, (b) performing an agglomerative hierarchical clustering based on an (n, n) array of distances between conformers, and (c) eliminating the structural redundancies via computation of an optimal number of clusters and selection of the conformers representative of each cluster.

Ligands Preparation. A database of 70 CYP 3A4 ligands (Figure 1) has been built by merging two ligands sets: one from a study of CYP 3A4 drug interactions[8] (34 compounds) and the other one from a quantitative structure−activity relationship (QSAR) study on CYP 3A4[9] (48 compounds). The interest of these data sets resides in the structural and size diversity of molecules, going from syn-Benzaldoxime Mw 121 to Cyclosporine Mw 1203, and in their various modes of action, recognized as either substrate or inhibitor of CYP 3A4. Each compound of this database has been built up and optimized under Sybyl 7.2[10] with conformational analysis achieved using the genetic algorithm implemented in Sybyl (GA conf search), or using 300−500 K molecular dynamics runs. Optimization was completed on the more stable conformer at the AM1/BCC level[11] using Gaussian 03 suite of programs,[12] despite that there is no need for high quality 3D structures serving as starting points for conformer generation.

Conformer Generation. The need to select diverse conformational ensembles for virtual screening has been

* Author to whom correspondence should be addressed. E-mail: michel.petitjean@cea.fr.
† CEA/DSV/iBiTec-S/SB[2]SM (CNRS URA 2096).
‡ UFR de Chimie, Université Louis Pasteur.

10.1021/ci800275k CCC: $40.75
Published on Web 01/30/2009

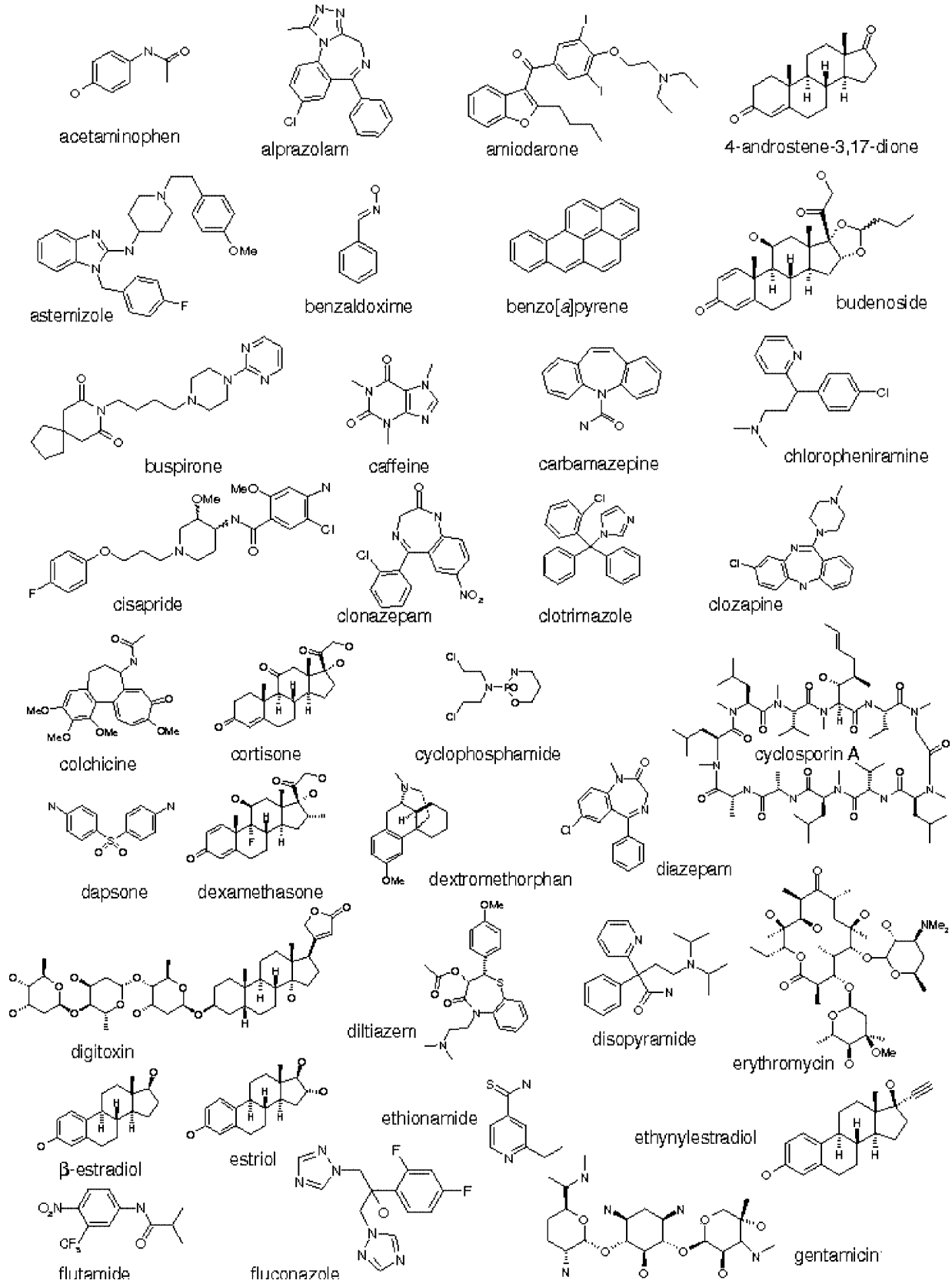

Figure 1. Part 1 of 2.

recently pointed out,[13] and the size variations of conformational ensembles needed to cover the conformational space have been intensively investigated by Borodina et al.[14] using Omega. The main differences between this latter approach and ours are the following: (i) the method of Borodina et al. does not involve clusters, and (ii) it is based upon an arbitrary root mean squared deviation (rmsd) resolution which is not needed in our present approach. In a review of conformers generators,[15] it has been concluded that geometrically similar structures should be collected in order to increase the probability of finding the bioactive conformation among the generated ensembles, but it has been proved difficult to

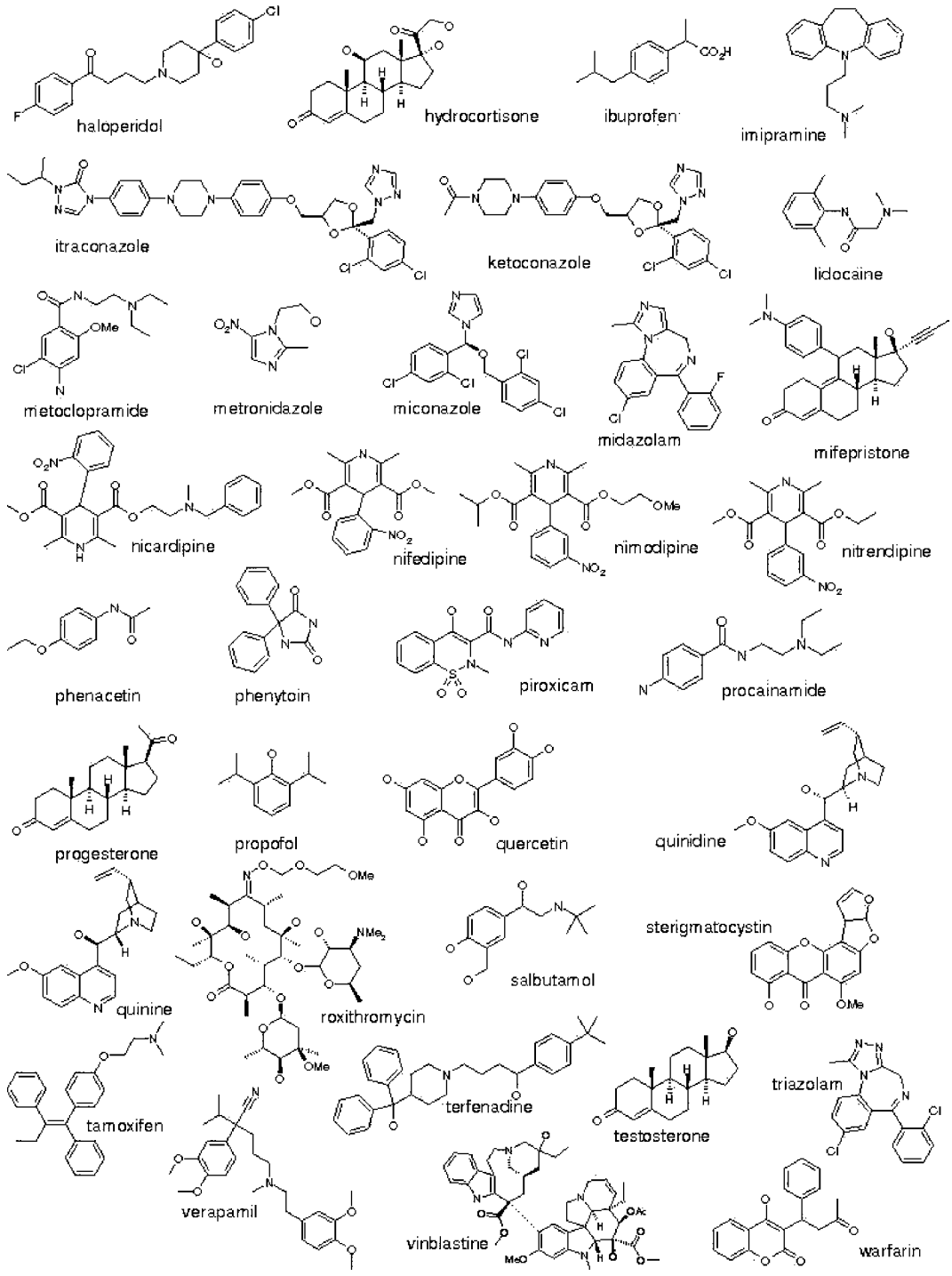

Figure 1. Part 2 of 2. Seventy CYP 3A4 test ligands.

retrieve bioactive structures having eight or more rotatable bonds for several conformational searching tools (Catalyst, Confort, Flo99, and Omega). There are some web access generators, such as Frog,[16] but this is limited to 100 generated conformers. The Sybyl software generator[17] handles stereochemistry constraints, offers both a systematic search and a random search to explore the conformational space, and has been retained in the framework of our study. The full conformer generation procedure we have built is callable under the Unix shell and does not require graphic inputs. It is constituted by SPL (Sybyl programming language) modules and Unix modules. We did not use the Sybyl genetic algorithm search because it is not callable in SPL in Sybyl 8.0.

Starting from some low energy initial conformation, the number N_{RB} of acyclic rotatable bonds is first computed by Sybyl. When $N_{RB} \leq 2$, the conformers generated by the random search are added to those generated by the systematic search with an increment angle of 30°. When $3 \leq N_{RB} \leq 4$, the conformers generated by the random search are added to those generated by the systematic search (up to 250 conformers) with an increment angle of 60°. When $N_{RB} > 4$, the conformers are generated solely by the random search. It is pointed out that the random search is useful to handle ring conformations, discarding the value of N_{RB}. In any case, the cutoff of the energy increase from each initial conformer was fixed to 50 kcal/mol to guarantee the geometric diversity.

At the end of the conformer generation step, it is expected to get a sufficient coverage of the conformational space, so that most probable conformations of the ligand are retrieved. However, the procedure above generates a high redundancy. Keeping a huge number of unnecessary conformations would lead to redhibitory computational times during the virtual screening step. The redundancy must be reduced, and only the conformers representative of the geometric conformational diversity of the ligand have to be retained.

Clustering. Let n be the number of conformers initially returned by the conformer generator for a given ligand. The rms distances between each pair of conformers are computed with the ARMS software,[18] which is based on a quaternion representation of rotations in order to ensure that only proper rotations are allowed during the spatial alignments.[19,20] ARMS returns also the maximal common 3D motif, using the SDM algorithm.[21,22]

The n by n symmetric matrix of rms distances is submitted to an ascending hierarchical clustering. Starting from n groups of one conformer, the two closest groups are merged and so on until it remains only one group of n conformers. At each of the $n - 1$ steps of the clustering, a score function is computed to evaluate the current set of clusters. The maximum of this score function is used to fix the final number K of clusters, and the hierarchical clustering is again performed but it is stopped when there are exactly K clusters. Then, for each cluster, a mean conformer representative of the cluster is computed.

Given an array of distances between conformers, there are several possible metrics to calculate the distance D between two groups of conformers:[23] when the groups j_1 (size n_1) and j_2 (size n_2) are merged, the metric defines how the distances are recomputed between the group j_1j_2 and the other groups. We have used the single linkage ($D(i, j_1j_2) = \min\{D(i, j_1), D(i, j_2)\}$), the complete linkage ($D(i, j_1j_2) = \max\{D(i, j_1), D(i, j_2)\}$), and the following average linkage: $D^2(i, j_1j_2) = [n_1D^2(i, j_1) + n_2D^2(i, j_2)]/(n_1 + n_2)$.

Finding an adequate stop criterion for the hierarchical clustering is a difficult problem. Many criterions are based on an arbitrary fixed critical distance threshold,[24,25] or calculate F as the ratio of the sample variance between clusters to that within clusters, and then calculate the probability of F to exceed a fixed probability level.[24] The optimal selection of such arbitrary values is difficult to do in a virtual screening context. Some other criterions can be found in recent books,[26,27] and more can be found in the specialized literature out of the chemistry field.[28−34] Methods vary widely in their performance,[35] and no clear conclusion arose from comparative studies.[29,31] We have discarded the methods based on some arbitrary fixed critical value, those only adequate for small size sets, and some of those assuming an euclidean distance. In this latter situation, it is recalled that our conformers are not located in the space: only their rms distances are known. The rms distance being not euclidean, we computed the inertias by formal analogy to the euclidean case (eq 5 in Appendix 1), and we used approximate mean points when needed (see Appendix 2). The stop criteria which were unable to reduce the conformational redundancy (i.e., returning most time n conformers, the stop failed), were not retained: Dunn,[34] Silhouette,[27] Davies and Bouldin,[32] Calinski and Harabasz,[28] Point biserial,[28] Stepsize,[28] and TraceW,[28] McClain and Rao,[28] and Hartigan.[28] When mean points are needed, they were computed as described in Appendix 2 because the space is not euclidean. Only a variant of the clustering gain criterion[36] has overcome all the limitations above.

The clustering gain G was originally defined for data in euclidean spaces. It is computable as follows: $G = \sum_{j=1}^{j=K} (n_j - 1)d^2(g_j, g)$, where n_j is the size of the cluster j, and $d(g_j, g)$ is the distance from its mean point to the global mean point of the full set of the n conformers. Following the notations of Appendix 1, $G = \sum_{j=1}^{j=K} (n_j - 1)Tr(B_j)/n_j$. Because we are not working in an euclidean space, we computed the mean points as described in Appendix 2. The clustering gain is a non-negative function of the number of clusters, taking the value zero at the beginning of the hierarchical clustering because all values n_j are equal to 1 and being again null at the end of the clustering because $g_j = g$. Thus G has indeed a maximum, even in the noneuclidean case. If it happens that there are several absolute maxima, the smallest number of clusters is retained. As an exception, when the clustering gain is null for all n values, the n conformers are retained. This latter situation, which is unlikely to occur for high n values, was indeed observed only once, for $n = 3$.

After computation of the clusters, the conformer representative of each cluster was selected such that it minimizes the quadratic mean of its rms distances to all members of its cluster. This quadratic mean is also a measure of the dispersion within the cluster (see Appendix 2).

3. RESULTS

Each optimized conformer was stored in a separate mol2 file and then was submitted to the Sybyl 8.0 conformer generator. The resulting sets of conformers were submitted to the hierarchical clustering with the clustering gain G as stop criterion. The number of generated conformers, the numbers of clusters, and the dispersions of the mean cluster points are reported in Table 1 for each of the three linkage methods (single, average, and complete). The dispersion of the mean cluster points is their quadratic mean distance (clusters sizes discarded) to the global mean point, all mean points being computed as described in Appendix 2. We observed that these dispersions are weakly dependent on the linkage method. The variations of the clustering gain as a function of the number of classes have been reported for some selected ligands in Figure 2. It is observed that the clustering gain function takes low values for high number of classes and then exhibits its maximum for low numbers of classes. Some secondary maxima appear but are not interpreted. Both the original clustering gain theory[36] and our modified clustering gain are based upon a global maximum.

As expected, the benzopyrene has been assigned only one representative conformer, but it is because Sybyl did not generated any conformer in addition to the input one. In the steroid series, progesterone and testosterone differ by one functional group and one rotatable bond (acetyl vs hydroxy on the carbon 17) and were assigned 13−14 conformers for the average and the complete linkage, although the single linkage lead respectively to 2 conformers and 21 conformers

Table 1. Properties of conformers generated by Sybyl[a]

ligand	n	K_s	K_a	K_c	σ_s	σ_a	σ_c
acetaminophen	276	22	8	11	1.19	1.16	1.19
alprazolam	76	5	4	4	1.51	1.65	1.65
amiodarone	377	65	8	10	3.16	2.93	3.02
4-androstene-3,17-dione	111	6	9	3	0.42	0.54	0.48
astemizole	237	27	18	19	3.37	3.39	3.39
benzaldoxime	9	14	4	8	0.80	0.75	0.81
benzo(a)pyrene	1	1	1	1	0.00	0.00	0.00
budenoside	78	8	4	12	1.78	1.74	1.71
buspirone	199	43	22	10	2.84	2.87	2.64
caffeine	39	4	5	4	0.75	0.74	0.70
carbamazepine	123	12	3	9	0.79	0.72	0.67
chlorpheniramine	108	18	8	10	2.29	2.24	2.21
cisapride	192	36	13	11	3.28	3.33	3.03
clonazepam	41	3	3	4	1.45	1.45	1.28
clotrimazole	24	5	6	4	2.01	2.07	1.95
clozapine	67	5	3	7	1.89	1.78	1.91
colchicine	130	3	2	4	2.48	2.39	2.31
cortisone	30	8	6	5	1.01	1.06	0.97
cyclophosphamide	240	43	10	3	2.02	1.97	1.83
cyclosporin	21	4	4	3	2.43	2.33	2.48
dapsone	275	8	8	8	1.86	1.86	1.86
dexamethasone	45	3	4	4	1.35	1.52	1.52
dextromethorphan	179	14	20	19	0.99	0.93	0.87
diazepam	40	5	6	6	1.59	1.49	1.49
digitoxine	34	8	5	4	3.30	3.24	3.15
diltiazem	188	6	5	11	2.98	2.98	2.80
disopyramide	279	95	3	5	2.72	2.61	2.51
erythromycin	31	11	8	5	2.38	2.39	1.83
estradiol	127	9	10	9	0.80	0.77	0.80
estriol	273	24	31	29	0.42	0.43	0.43
ethionamide	296	22	14	17	1.40	1.37	1.29
ethynylestradiol	3	1	1	1	0.00	0.00	0.00
fluconazole	181	7	3	7	2.26	2.26	2.37
flutamide	7	9	8	8	1.91	1.85	1.88
gentamycin	17	22	6	5	3.12	3.06	2.74
haloperidol	218	25	8	5	2.74	2.67	2.77
hydrocortisone	32	4	5	5	1.08	1.06	1.06
ibuprofene	41	2	5	2	1.72	1.85	1.83
imipramide	144	11	5	8	2.37	2.25	2.26
itraconazole	199	21	4	9	3.94	3.92	3.56
ketoconazole	240	17	3	10	3.35	3.21	3.27
lidocaine	196	33	4	16	2.47	2.34	2.30
metoclopramide	375	21	10	2	2.55	2.49	2.50
metronidazole	9	2	2	2	1.32	1.32	1.32
miconazole	231	18	4	9	2.73	2.59	2.55
midazolam	52	4	3	3	0.94	1.04	1.04
mifepristone	2	1	1	1	0.00	0.00	0.00
nicardipine	196	7	10	3	3.39	3.12	3.50
nifedipine	112	4	4	4	2.00	2.00	2.00
nimodipine	232	43	12	4	2.76	2.78	2.79
nitrendipine	134	12	5	14	2.39	2.27	2.33
phenacetin	33	9	9	8	1.52	1.51	1.45
phenytoin	11	2	2	3	1.22	1.22	1.43
piroxicam	48	5	8	5	2.53	2.62	2.30
procainamide	285	25	9	2	2.38	2.36	2.40
progesterone	42	2	14	13	0.71	0.71	0.71
propofol	7	4	2	3	1.75	1.25	1.46
quercetin	5	2	2	2	1.11	1.11	1.11
quinidine	107	7	6	4	2.41	2.38	2.33
quinine	124	2	2	5	2.53	2.53	2.45
roxithromycin	45	12	6	3	3.00	2.99	2.92
salbutamol	171	25	3	5	2.25	2.49	2.21
sterigmatocystin	83	4	3	3	0.48	0.59	0.57
tamoxifen	359	56	10	5	3.06	2.82	2.69
terfenadine	253	79	8	4	3.40	3.37	3.36
testosterone	103	21	13	14	0.61	0.66	0.65
triazolam	52	3	3	3	1.11	1.11	1.11
verapamil	339	78	13	6	3.48	3.39	3.48
vinblastine	36	2	3	3	3.14	2.69	2.69
warfarin	86	20	7	8	2.25	2.16	2.14

[a] n: number of conformers generated by Sybyl. Numbers of representative conformers: K_s (single linkage), K_a (average linkage), K_c (complete linkage). Dispersions of clusters, in angstroms: σ_s (single linkage), σ_a (average linkage), σ_c (complete linkage).

(Sybyl generated 42 conformers for progesterone vs 103 for testosterone). It is assumed that the acetyl group of progesterone collides its spatially neighboring methyl group (carbon 18), thus lowering the number of conformers. All linkage methods lead to attribute more representative conformers for estriol (24−31 conformers) than for estradiol (9−10 conformers), although they differ only by the hydroxy group on the carbon in position 13. This hydroxy group permitted to multiply roughly by 2−3 the number of conformations (one more rotatable bond). Sybyl found only three conformers for ethynylestradiol. In fact, the clustering gain was null for all values of the number of classes, i.e. from 1 to 3. It can be noticed that the ethynyl group induces steric constraints in the region of the quaternary carbons 13 and 17, such that it could explain why Sybyl generated few conformers. The same kind of remark applies to mifepristone, for which Sybyl generated only two conformers.

The macrocyclic antibiotics (cyclosporin, erythromycin, roxythromycin) offer rather few representative conformations, particularly for the average and the complete linkage. Cyclosporin, which is by far the largest of the three macrocycles, was assigned four conformations for the single and the average linkage and only three for the complete linkage (Sybyl generated 20 conformers). The same three representative conformers were retrieved by all linkage methods. From a steric point of view, cyclosporin is the largest CYP 3A4 substrate in our database. In this regard, information about its possible conformations along its trajectory in the channels leading to the active site of the CYP 3A4 are of high value to characterize these channels. Simulation studies[37] of cyclosporin A indicate that it could exist several conformations of similar stability to those of the two known experimental conformations. A study in organic solvent[38] proved that there are three experimental conformations rather than two. There is thus an agreement between our results and those of the literature regarding the existence of three main conformations.

As a whole, large numbers of representative conformers have been generated for all linkage methods for ligands exhibiting significant acyclic chains, and particularly for the single linkage. There is no clear trend in the linkage methods when going from single to average and to complete linkage: e.g., the respective number of representative conformers for ethionamide and for clonazepam are respectively 22, 3, 17, and 3, 3, 4 for the three linkage methods. This tells us that the effect of the linkage method depends on the data and that it is difficult to tell which linkage is the most adequate.

4. DISCUSSION AND CONCLUSION

It must be emphazised that an infinity of conformers is theoretically allowed to exist in the conformational space, and thus, there is no universal definition of the number of *representative* conformers, even in the framework of our particular study. So, having different numbers of clusters for a given chemical at the end of each clustering experiment is not surprising, and setting some criteria for definitely ranking the three linkage methods would be highly subjective. It is pointed out that more variants of the clustering methods are possible and that, for any of them, the final results are dependent on their input data, i.e. the conformer generator. When the conformer generator uses a Monte-Carlo method,

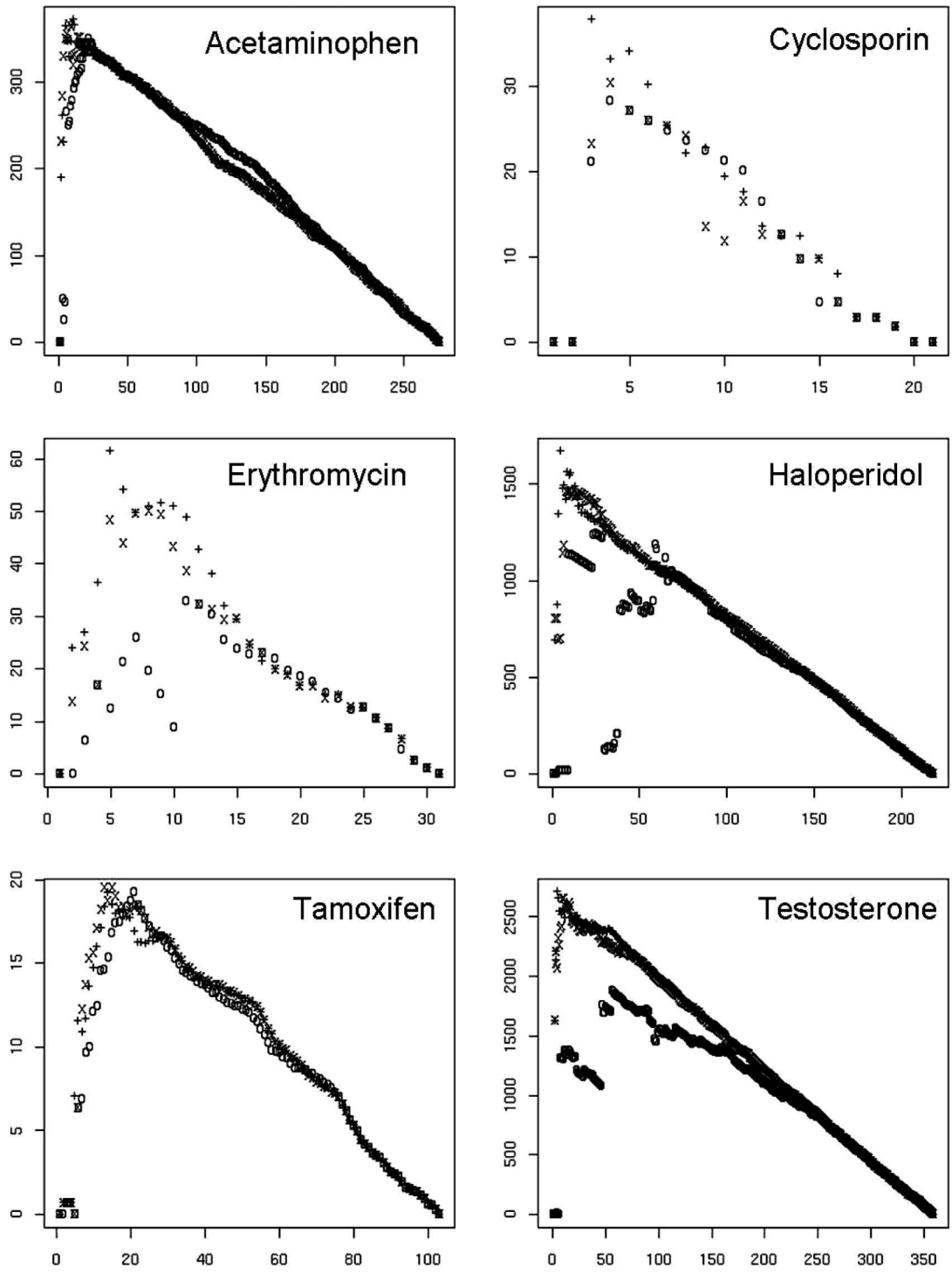

Figure 2. Variations of the clustering gain as a function of the number of classes for some selected ligands: (○) single linkage; (×) average linkage; (+) complete linkage.

it may be programmed to generate non-reproducible random sequences, so that even the results of the full procedure may fluctuate. This is the case with Sybyl, although it has been checked at several occasions that the final numbers of conformers vary weakly within each linkage method.

This means that an advanced study involving the comparison of results for various chemical libraries, for several conformer generators, and for more clustering methods, would imply a combinatorial comparison of a huge number of results which is outside the scope of this study. However, only the modified clustering gain stop criterion that we have defined here was able to provide acceptable numbers of conformations for all members of our CYP 3A4 ligands database. For most compounds, other stop criterions exhibit their optimum value either at the first step or at the last step of the clustering, indicating that, in fact, the stop failed. In these situations, the resulting numbers of conformers were not accepted because they were computed with a stop criterion which visibly does not work with our data, even if the final result seems satisfactory in some situations (e.g., only one conformer for the benzopyrene). The stop criterion must work for all compounds: we cannot prove that it works in all situations, but we can detect that it fails in several situations. This failure never occurred for the modified

clustering gain, whereas it occurred for the other stop criteria. The number of conformers is expected to increase with the number of rotatable bonds. On the other hand, the rotatable bonds should be handled with caution: e.g., the situation of a methyl group is different from the situation of a *t*-butyl group due to steric hindrance. Counting rotatable bonds in rings and evaluating their impact upon the number of conformers to generate is difficult: e.g. we can flag six rotatable C−C bonds in the cyclohexane by a symmetry argument, but most chemists would like to retain only two major conformations. The size of the molecule, the number of heavy atoms, and the number of double and triple bonds must be taken in account with care, too. A full discussion of the number of conformers is outside the scope of this paper.

There may be several conformations along the trajectory from the exterior of the enzyme to its active site, which may differ from the bioactive ones in the final enzyme−ligand complexes and from those of the X-ray structures. This is why we have neither reported the energies nor the rmsd from the X-ray structures: these data are difficult to use in our context. Finally, validating experimentally the efficiency of our selection process is a non-trivial task, and this point will be considered in a further study, in which further evaluation of the method will be carried out.

ACKNOWLEDGMENT

We thank the three reviewers for having carefully read our manuscript and for their useful suggestions.

APPENDIX 1: COMPUTING INERTIAS

We consider n points in R^d, and we note g their barycenter. Let x_i ($i = 1, 2,..., n$) be the respective vectors originating at g associated to these n points. The set of the n points is partitioned into K clusters, and the vectors associated to the points in the cluster j are noted x_i^j ($i = 1, 2,..., n_j$), n_j being the number of points in the cluster j ($j = 1, 2,..., K$). We associate the mean point of the cluster j to the vector $g_j = (\sum_{i=1}^{i=n_j} x_i^j)/n_j$.

The quote denoting a matrix transposition, the contribution W_j of the cluster j to the within clusters inertia matrix W, is the following:

$$W_j = \sum_{i=1}^{i=n_j} (x_i^j - g_j)(x_i^j - g_j)' \quad (1)$$

$$W = \sum_{j=1}^{j=K} W_j \quad (2)$$

The contribution B_j of the cluster j to the between clusters inertia matrix B is the following:

$$B_j = n_j(g_j - g)(g_j - g)' \quad (3)$$

$$B = \sum_{j=1}^{j=K} B_j \quad (4)$$

The contribution T_j of the cluster j to the full inertia matrix T of the n points is $T_j = W_j + B_j$ and of course $T = W + B$. The inertias are the respective traces of the inertia matrices.

It can be easily checked that merging the clusters j_1 and j_2 decreases the between clusters inertia from the quantity $[n_{j_1}n_{j_2}/(n_{j_1} + n_{j_2})]d^2(g_{j_1} - g_{j_2}, g_{j_1} - g_{j_2})$, where d is the usual euclidean distance. The total inertia being invariant, it follows that the within clusters inertia increases from the quantity above, which is always non-negative.

The computation of the inertia of the cluster j can be performed only from the distances within the cluster, since we have the following:

$$Tr(W_j) = \left[\sum_{i_1=1}^{i_1=n_j} \sum_{i_2=1}^{i_2=n_j} d^2(x_{i_1}^j, x_{i_2}^j)\right] \Big/ 2n_j \quad (5)$$

Thus we can compute the inertia without the knowledge of any mean point, but it involves a double summation rather than a single summation. The same remark applies to any subset of the n points.

APPENDIX 2: DISTANCES TO THE MEAN POINT

The notations are those of Appendix 1, but we suppress any cluster index for clarity, and what follows works for any set of n points $x_1, x_2,..., x_n$ in R^d.

As is well-known, the barycenter g of the n points minimizes the quantity Δ:

$$\Delta = \min_{\{y \in R^d\}} \sum_{i=1}^{i=n} d^2(y, x_i)$$

Now, we perform the minimization above on a subset of R^d which is the set of the n points itself:

$$\Delta^* = \min_{\{y \in \{x_1, x_2,..., x_n\}\}} \sum_{i=1}^{i=n} d^2(y, x_i)$$

This minimum is reached for some point $y = x_k$ ($1 \le k \le n$), and $\Delta^* \ge \Delta$. Then, x_k is the best approximation of the barycenter of the n points among the n points themselves, i.e. it is the point of the set which is the closest to the barycenter of the set. The proof follows.

$$\sum_{i=1}^{i=n} d^2(x_j, x_i) = \sum_{i=1}^{i=n} d^2(x_i, g) + nd^2(x_j, g) = Tr(T) + nd^2(x_j, g)$$

$$\text{Thus, } d^2(x_j, g) = \left[\sum_{i=1}^{i=n} d^2(x_j, x_i)\right] \Big/ n - Tr(T)/n, \text{ so that}$$

$$\min_{\{y \in \{x_1, x_2,..., x_n\}\}} d^2(x_j, g) = \left[\min_{\{y \in \{x_1, x_2,..., x_n\}\}} \left[\sum_{i=1}^{i=n} d^2(x_j, x_i)\right]\right] \Big/ n - Tr(T)/n$$

The error is the following: $d^2(x_k, g) = [\sum_{i=1}^{i=n} d^2(x_k, x_i) - Tr(T)]/n$. It is null when it happens that the barycenter falls within the set of the n points. The calculations above extend trivially to the continuum of points and distributions, each summation symbol being replaced by the appropriate expectation operator.

REFERENCES AND NOTES

(1) Guengerich, F. P. Human Cytochrome P450 Enzymes. In *Cytochrome P450, Structure, Mechanism, and Biochemistry*, third ed.; Ortiz de Montellano, P. R., Ed.; Kluwer/Plenum: New York, 2005; Chapter *10*, section 6.20.3, pp 425−426.

(2) Lewis, D. F. V. Criteria Governing Substrate Selectivity for Human Hepatic P450s. In *Guide to Cytochromes P450, Structure and Function*; Taylor & Francis: London, UK, 2001; Chapter *4.4*, pp 102−109.

(3) Nguyen, T. A.; Tychopoulos, M.; Bichat, F.; Zimmermann, C.; Flinois, J. P.; Diry, M.; Ahlberg, E.; Delaforge, M.; Corcos, L.; Beaune, P.; Dansette, P.; André, F.; de Waziers, I. Improvement of Cyclophosphamide Activation by CYP2B6 Mutants: from in Silico to ex Vivo. *Mol. Pharmacol.* **2008**, *73*, 1122–1133.

(4) Lafite, P.; André, F.; Zeldin, D. C.; Dansette, P. M.; Mansuy, D. Unusual Regioselectivity and Active Site Topology of Human Cytochrome P450 2J2. *Biochemistry* **2007**, *46*, 10237–10247.
(5) Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-Based Models for the Isoform Specificity of Cytochrome P450 3A4, 2D6, and 2C9 Substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.
(6) Good, A. C.; Cheney, D. L. Analysis and Optimization of Structure-Based Virtual Screening Protocols (1): Exploration of Ligand Conformational Sampling Techniques. *J. Mol. Gr. Model.* **2003**, *22*, 23–30.
(7) Montes, M.; Braud, E.; Miteva, M. A.; Goddard, M.-L.; Mondésert, O.; Kolb, S.; Brun, M.-P.; Ducommun, B.; Garbay, C.; Villoutreix, B. O. Receptor-Based Virtual Ligand Screening for the Identification of Novel CDC25 Phosphatase Inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 157–165.
(8) Kenworthy, K. E.; Bloomer, J. C.; Clarke, S. E.; Houston, J. B. A. CYP 3A4 drug interactions: correlation of 10 in vitro probe substrates. *Br. J. Clin. Pharmacol.* **1999**, *48*, 716–727.
(9) Lill, M. A.; Dobler, M.; Vedani, A. Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. *ChemMedChem* **2006**, *1*, 73–81.
(10) *SYBYL 7.2*; Tripos International: St Louis, Mo, 2006.
(11) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(12) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; revision C.02, Gaussian, Inc.; Wallingford, CT, 2004.
(13) Loriot, S.; Sachdeva, S.; Bastard, K.; Prevost, C.; Cazals, F. *On the Characterization and Selection of Diverse Conformational Ensembles*; INRIA Research Report 6503, April **2008**; https://hal.inria.fr/inria-00252046 (accessed Sep 16, 2008).
(14) Borodina, Y. V.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space. *J. Chem. Inf. Model.* **2007**, *47*, 1428–1437.
(15) Boström, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137–1152.
(16) Bohme Leite, T.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tufféry, P. Frog: a FRee Online druG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35*, 568–572.
(17) Ghose, A. K.; Jaeger, E. P.; Kowalczyk, P. J.; Peterson, M. L.; Treasurywala, A. M. Conformational Searching Methods for Small Molecules. I. Study of the SYBYL SEARCH Method. *J. Comput. Chem.* **1993**, *14*, 1050–1065.
(18) Petitjean, M. http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html#ARMS (accessed Sep 16, 2008).
(19) Petitjean, M. On the Root Mean Square Quantitative Chirality and Quantitative Symmetry Measures. *J. Math. Phys.* **1999**, *40*, 4587–4595; see the Appendix.
(20) Petitjean, M. Chiral Mixtures. *J. Math. Phys.* **2002**, *43*, 4147–4157; Appendix A.5.
(21) Petitjean, M. Three-Dimensional Pattern Recognition from Molecular Distance Minimization. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1038–1049.
(22) Petitjean, M. Interactive Maximal Common 3D Substructure Searching with the Combined SDM/RMS Algorithm. *Comp. Chem.* **1998**, *22*, 463–465.
(23) Nakache, P.; Confais, J. Algorithmes d'agrégation fondés sur un lien métrique. In *Approche Pragmatique de la Classification*; Technip: Paris, France, 2005; Chapter *1.8*, pp 35−43.
(24) Shenkin, P. S.; McDonald, D. Q. Cluster Analysis of Molecular Conformations. *J. Comput. Chem.* **1994**, *15*, 899–916.
(25) Vesterman, B.; Golender, V.; Golender, L.; Fuchs, B. Conformer Clustering Algorithm and its Application for Crown-Type Macrocycles. *J. Mol. Struct.* **1996**, *368*, 145–151.
(26) Mirkin, B. Validity and Reliability. In *Clustering for Data Mining, A Data Recovery Approach*; Chapman & Hall/CRC: Boca Raton, FL, 2005; Chapter *7.5*, pp 232−243.
(27) Nakache, P.; Confais, J. Nombre de classes a retenir. In *Approche Pragmatique de la Classification*; Technip: Paris, France, 2005; Chapter *7*, pp 189−205.
(28) Milligan, G. W.; Cooper, M. C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **1985**, *50*, 159–179.
(29) Dubes, R. C. How Many Clusters are Best? - An Experiment. *Pattern Rec.* **1987**, *20*, 645–663.
(30) Jain, A. K.; Dubes, R. C. Cluster Validity. In *Algorithms for Clustering Data*; Prentice Hall: Englewood Cliffs, NJ, 1988; Chapter *4*, pp 143−222.
(31) Atlas, R. S.; Overall, J. E. Comparative Evaluation of Two Superior Stopping Rules for Hierarchical Cluster Analysis. *Psychometrika* **1994**, *59*, 581–591.
(32) Bezdek, J. C.; Pal, N. R. Some New Indexes of Cluster Validity. *IEEE Trans. Syst. Man Cyber.* **1998**, *28*, 301–315.
(33) Günter, S.; Bunke, H. Validation Indices for Graph Clustering. *Pattern Rec. Lett.* **2003**, *24*, 1107–1113.
(34) Bolshakova, N.; Azuaje, F. Cluster Validation techniques for Genome Expression Data. *Signal Process.* **2003**, *83*, 825–833.
(35) Mojena, R. Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *Comput. J.* **1977**, *20*, 359–363.
(36) Jung, Y.; Park, H.; Du, D.-Z.; Drake, B. L. A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. *J. Global Optim.* **2006**, *25*, 91–111.
(37) O'Donohue, M. F.; Burgess, A. W.; Walkinshaw, M. D.; Treutlein, H. R. Modeling Conformational Changes in Cyclosporin A. *Protein Sci.* **1995**, *4*, 2191–2202.
(38) Verheyden, P.; Jaspers, H.; De Wolf, E.; van Binst, G. Conformational Study of Cyclosporin A in Acetone at Low Temperature. *Int. J. Pept. Protein Res.* **1994**, *44*, 364–371.

CI800275K