BIOINFORMATICS

Vol. 22 no. 14 2006, pages e58–e65
doi:10.1093/bioinformatics/btl212

ACIAP, Autonomous hierarchical agglomerative Cluster Analysis based protocol to partition conformational datasets

Giovanni Bottegoni[1], Walter Rocchia[2], Maurizio Recanatini[1] and Andrea Cavalli[1,*]

[1]Department of Pharmaceutical Sciences, University of Bologna, Via Belmeloro 6, I-40126, Bologna, Italy and [2]NEST CNR-INFH, Scuola Normale Superiore of Pisa, Piazza dei Cavalieri, 7, I-56126, Pisa, Italy

ABSTRACT

Motivation: Sampling the conformational space is a fundamental step for both ligand- and structure-based drug design. However, the rational organization of different molecular conformations still remains a challenge. In fact, for drug design applications, the sampling process provides a redundant conformation set whose thorough analysis can be intensive, or even prohibitive. We propose a statistical approach based on cluster analysis aimed at rationalizing the output of methods such as Monte Carlo, genetic, and reconstruction algorithms. Although some software already implements clustering procedures, at present, a universally accepted protocol is still missing.

Results: We integrated hierarchical agglomerative cluster analysis with a clusterability assessment method and a user independent cutting rule, to form a global protocol that we implemented in a MATLAB metalanguage program (ACIAP). We tested it on the conformational space of a quite diverse set of drugs generated via Metropolis Monte Carlo simulation, and on the poses we obtained by reiterated docking runs performed by four widespread programs. In our tests, ACIAP proved to remarkably reduce the dimensionality of the original datasets at a negligible computational cost. Moreover, when applied to the outcomes of many docking programs together, it was able to point to the crystallographic pose.

Availability: ACIAP is available at the ''ACIAP'' section of the website http://www.scfarm.unibo.it.

Contact: E-mail: andrea.cavalli@unibo.it.

Supplementary Information: The complete series of ACIAP results is available in the ''services'' section of the website http://www.scfarm.unibo.it.

1 INTRODUCTION

The physicochemical and biological properties of a molecule critically depend upon conformations the molecule can adopt. Therefore, carrying out exhaustive and meaningful conformational analysis is pivotal for deeply investigating any molecular feature. For instance, any three-dimensional ligand-based approach in drug design can't help using a complete analysis of the conformational space. Monte Carlo simulation is just one of the methods available to achieve this sampling (Chang, *et al*., 1989). In a Monte Carlo study, the conformational space of a molecule is sampled by randomly changing dihedral angle rotations or atom Cartesian coordinates. If the currently drawn sample is lower in energy than its predecessor, then it is retained as a starting point for the successive iteration. Conversely, when the new conformation is higher in energy, it can be retained according to two alternative criteria: either its energy belongs to a predefined window or the ''move'' can be accepted with a probability related to the Boltzmann factor, following the Metropolis method (Metropolis, *et al*., 1953).

Two fields that make a great use of conformational sampling are docking and virtual screening, both of them holding a prominent position in the modern structure-based drug design (Taylor, *et al*., 2002). In a limited computational time, they have to face a hard two-fold problem: generating a sensible conformational ensemble and then ranking its members. Besides Monte Carlo sampling, the ligand conformational space can be explored by genetic and incremental algorithms.

Apparently, sampling is an easier job to do than scoring. In fact, reiterated docking runs usually provide at least one pose close to the crystallographic one. In contrast, due to different heuristics and approximation levels, scoring functions do not always succeed in including the crystallographic pose among the most favorable ones. On top of it, it is not unusual to see quite different rankings by some among the most widespread docking tools. In general, it cannot be said that one method outperforms the others, since different target and compound classes can lead to different performances. A number of different possibilities rather than a single binding mode can be obtained also as a result of reiterated runs of the same algorithm, when it adopts a random based approach. Due to the computational cost of the sampling process and of the evaluation of the binding free energy, it would be definitely useful to have a restricted, but still representative, set of conformations to be processed with more thorough techniques.

Cluster Analysis (CA) is a discipline that encompasses a number of different algorithms to partition samples in homogeneous classes without any *a priori* knowledge. It is already used to analyze the large amount of data generated by molecular modeling software, such as the outcomes of conformational analysis and docking outputs (Chema, *et al*., 2004).

In principle, there does not exist a unique ''correct'' method to cluster a dataset, and a large number of variations have been devised, from which one has to choose the most appropriate one.

As an example, X-cluster, developed in 1994 by Shenkin and McDonald (Shenkin and McDonald, 1994) and implemented in the MacroModel software package (Mohamadi, *et al*., 1990) is one of the most widely exploited algorithms for organizing the output of conformational sampling. X-cluster employs a hierarchical agglomerative approach with the single linkage rule (see the Algorithm Description section for further details). As a major drawback for any

*To whom correspondence should be addressed.

automated procedure, X-cluster leaves to the user the choice of the most suitable clustering level.

In docking and virtual screening simulations, some programs (such as AutoDock and GOLD) implement CA to better rationalize their outcomes. In particular, AutoDock sorts conformations by increasing energy and then implements a nonhierarchical clustering method with single linkage rule to partition the poses. The clustering process always starts from the best scoring pose, and, due to the peculiarity of the single linkage rule, first clusters tend to be the more numerous. The process is iterated through conformations, grouping together the elements whose Root Mean Square Deviations (RMSDs) are within a user-defined threshold value. In turn, GOLD has a dedicated utility (rms_analysis) to perform CA on the docked poses with a hierarchical agglomerative approach based on the complete linkage rule. This is known to be a non space-conservative linkage criterion that tends to create compact clusters of similar dimension. Moreover, no cutting rule is implemented in the CA of the GOLD program.

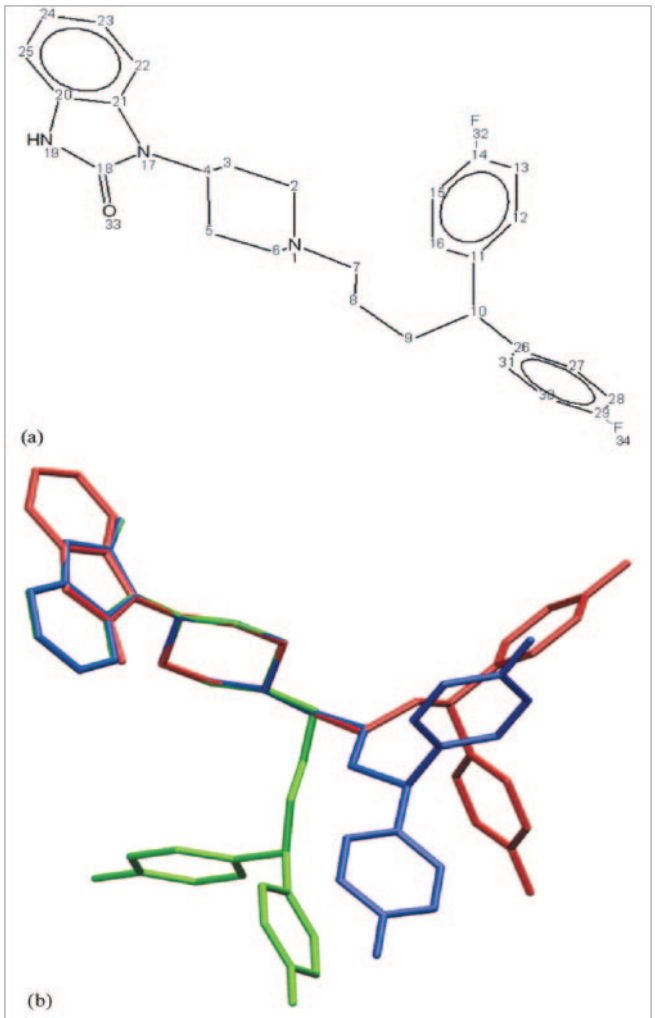
A suitable CA protocol should be able to provide a functional classification, i.e., to identify few conformations worthy to be further studied. Moreover, the protocol should be ''information'' driven and should not, in general, necessitate of any preexisting knowledge about the specificities of the target. Recently, we carried out a comparative study (Bottegoni, *et al*., 2006) about the use of different hierarchical agglomerative clustering rules associated with a user-independent cutting function applied to the outcomes of four different docking programs. From that study, we learned that the combination of an *a priori* clusterability assessment with the average linkage rule, and with a stopping criterion based on the Kelley-Gardner-Sutcliffe (KGS) penalty function (Kelley, *et al*., 1997) provides a good basis to achieve a sensible partitioning of conformational datasets.

In this work, we describe the implementation of our novel protocol in a MATLAB (The MathWorks, Inc.) metalanguage program, named AClAP (Autonomous hierarchical agglomerative Cluster Analysis based Protocol), and we discuss its performance vs. commonly available CA-based methods. AClAP design benefits from the understanding we gained from a conformational analysis we made over a set of ten marketed drugs with the aid of MacroModel (Mohamadi, *et al*., 1990) and over the above mentioned docking results, which concern a quite diverse set of ligands co-crystallized with different biological counterparts. Docking simulations were carried out by means of four programs, namely Dock (Ewing, *et al*., 2001), AutoDock (Morris, *et al*., 1998), GOLD (Jones, *et al*., 1997), and FlexX (Rarey, *et al*., 1996). Moreover, we statistically analyze the whole set of obtained conformations, and finally we discuss the behavior of the KGS penalty function.

Summarizing, AClAP turned out to meet all of the criteria required for a robust clustering protocol at a very limited computational cost. Therefore, we propose it as an innovative and user-friendly tool, which can be of great help to molecular modelers dealing with both ligand- and target-based drug design.

2 METHODS

AClAP is an interactive MATLAB metalanguage program that can take data from the widespread mol2 file format. AClAP can also take in input the torsion angles either in raw or csv (comma separated values) formats. It is able to automatically identify the number of poses and the set of nonhydrogen atoms.



Torsions #	1	2	3	4	5	6	7
Atoms	18-17-4-3	6-1-7-8	1-7-8-9	7-8-9-10	8-9-10-11	9-10-11-12	9-10-26-27
Conf. #1 (blue)	115.93	63.56	58.76	173.89	66.55	-122.36	-54.14
Conf. #2 (green)	116.09	170.48	-58.76	-173.83	166.51	-145.90	-56.72
...							
Conf. #121 (red)	-61.92	68.42	-106.91	-70.16	166.74	-144.22	-55.90

Fig. 1. Example of construction of matrix **M** for Pimozide. **a**) The nonhydrogen atom numbering and the acyclic torsion angles are reported. **b**) Pimozide chemical structure. The parameters reported in the table give rise to *n* by *d* matrix **M**, where *n* is the number of the sampled conformations (in the present example, 121) and *d* is the number of degrees of freedom of the molecule (in the present example, the 7 acyclic torsion angles of Pimozide).

Each conformation is considered as an observation in a *d*-dimensional space, and it is stored in an *n* by *d* matrix **M**, where *n* is the number of the sampled conformations and *d* is the number of degrees of freedom of the molecule (Figure 1). These latter include dihedral angle values of all rotatable bonds and the Cartesian coordinates of three atoms (limited to the clustering of docking poses), which account for global rotation and translation. Each column of **M** is z-standardized for subsequent

```
ACLAP - Version 1.0


"Data, data everywhere...."


Perform H* Test ? [y/n]
Select: y
Are poses superimposed ? [y/n]
Select: y
Poses file: prazosina_fit.mol2

Heavy Atoms Number   28

Parsing poses, please wait....


Poses Detected   24

Torsions file: prazosin.csv
Torsion File Format ? [csv - vector]
Select: csv
More poses to cluster ? [y/n]
Select: n
H* value is :   0.87

Procede to Cluster Analysis ? [y/n]
Select: y
Compare to a Reference Structure ? [y/n]
Select: n
Define symmetric atomic couples ? [y/n]
Select: n

Linkage Method:
[average - single - ward]
Seleziona: average
Plot the KGS Penalty Score Function ? [y/n]
Select: y
Save the current plot ? [y/n]
Select: n

Minimum KGS Penalty Function:     16.9
N. of Clusters:      5
N. of non-singleton Clusters:   5
Most Populated Cluster is n. 2 (10 elements)

Do you want to write a Report output ? [y/n]
Select: y
Report File name: prazosin_rep
```

Fig. 2. The dialog box of AClAP.

processing. The **M** matrix is exploited within the clusterability assessment, whereas, for CA, the full Cartesian coordinate set is used. In Figure 2, a typical dialog box of AClAP is shown.

2.1 Clusterability assessment

To assess whether conformations show a natural tendency to group into clusters, we implemented a modified version of a test originally developed by Hopkins (Hopkins, 1954): the H^* test.

This test is aimed at distinguishing between three main possibilities for the distribution of the members of the dataset: uniformly scattered, regularly spaced or naturally grouping. Only in the last case, CA is really justified. The H^* test is implemented as follows: first, a Principal Component Analysis is performed over the z-standardized matrix **M**; in order to lower the dimensionality of the problem, the original dataset is projected onto the reduced space L induced by the first three principal components. Then, a small number s of random points in L is generated. These points are normally distributed, with zero means, and their projection over each principal component direction has the same standard deviation as the corresponding principal component of the dataset. In our test, $s = n/20$. Now, s poses are randomly drawn and for each of them, as well as for each random point, the minimum distance to the members of the dataset is calculated, and named D_i for the poses, and V_i for the points. This procedure is repeated n times and the H^* value is calculated as the following average:

$$H^* = \left\langle \sum_{i=1}^{s} V_i \Big/ \left(\sum_{i=1}^{s} V_i + \sum_{i=1}^{s} D_i \right) \right\rangle_{dataset}, \quad (1)$$

Three cases can occur:

$0.5 \leq H^* \leq 0.6$ the poses are homogenously distributed
$H^* \to 0$ the poses are regularly spaced
$H^* \to 1$ the poses show a natural tendency to cluster

A cluster analysis should be carried out only in the last one. The absence of regular or repetitive patterns in the outcomes of conformational analysis and docking simulations makes unlikely the occurrence of the second case.

2.2 Cluster Analysis

AClAP implements a hierarchical agglomerative clustering algorithm. ''Hierarchical'' means that clusters at a higher level are union of clusters at lower levels, while ''agglomerative'' means that clusters never break apart during the formation process. The global hierarchy can be represented by means of a dendrogram, a tree showing different clustering levels, spanning from 1 to n. RMSD is taken as a measure of conformation-to-conformation distance. Therefore, the clustering algorithm starts with n unary clusters; at each step, the two closest clusters are merged, until only one cluster containing all the poses is reached. The way the inter-cluster distance is evaluated is called linkage rule. In AClAP, we implemented three among the most widely used linkage rules: single linkage, average linkage, and the Ward method. Single linkage (Everitt, *et al*., 2001), also known as nearest-neighbor distance method, defines distance as the one of the closest pair of conformations:

$$\Delta_{M,Q} = \min_{m \in \{1,\ldots,\chi_M\}, q \in \{1,\ldots,\chi_Q\}} (d_{m,q}), \quad (2)$$

where uppercase roman letters indicate clusters, d is the RMSD-based conformation distance, Δ is the inter-cluster distance, χ is the cardinality of a cluster.

A well-known drawback of single linkage rule is the so-called ''chaining'' phenomenon: first clusters naturally tend to incorporate the nearby conformations, therefore forming a ''chain''; as a consequence, there is a strong bias towards the first clusters to being more populated than others.

In the average linkage (Everitt, *et al*., 2001) method, the mean distance between all pairs of conformations is taken:

$$\Delta_{M,Q} = \frac{1}{\chi_M \chi_Q} \sum_{m=1}^{\chi_M} \sum_{q=1}^{\chi_Q} d_{m,q}. \quad (3)$$

According to this definition no conformation/cluster is preferred with respect to the others, preventing ''chaining'' effect to occur.

Finally, in AClAP, the Ward method can also be selected. This method uses a distance definition based on the analysis of variance (Ward and Hook, 1963). It attempts to minimize the Sum of Squares of any two potential clusters that can be formed at each step. This method tends to create a consistent number of small clusters. Our previous comparative study (Bottegoni, *et al*., 2006) led us to prefer the average linkage rule with respect to both single linkage and the Ward method.

When clustering is finished, the complete dendrogram is obtained and, for each cluster at each level, the so-called centroid can be calculated. The centroid is a ''hypothetical'' conformer whose coordinates are the average coordinates of all the cluster members. The representative conformer for a cluster is chosen as the conformation closest to the centroid. If the homogeneity requirement for the current cluster is fulfilled, the choice of the representative conformer is not expected to be critical.

2.3 Cutting rule

Once the dendrogram is formed, the crucial decision is to fix the level of clustering more suitable to represent the conformational space of interest. As it is natural for a hierarchical agglomerative approach, a tradeoff must be found between the overall number of clusters and the diversity among the conformations that belong to each cluster. AClAP adopts the KGS penalty function. The method is thoroughly described in the paper by Kelley *et al*. (1997) and here summarized and discussed (see Results and Discussion for further details).

An average spread value is calculated for each clustering level of the dendrogram, for simplicity of representation, it is numbered with respect to the number of clusters of the level:

$$AvS_w = \frac{1}{w}\sum_{M=1}^{w} S_M, \tag{4}$$

where w is the number of clusters at a fixed clustering level and S_M is the spread of the M-th cluster, defined as follows:

$$S_M = \frac{2}{\chi_M(\chi_M - 1)}\sum_{m=1}^{\chi_M}\sum_{q=m+1}^{\chi_M} d_{m,q}. \tag{5}$$

When all average spread values are collected, they need to be normalized so that they lie between 1 and $n-1$. The penalty P_w is therefore calculated as:

$$P_w = \frac{(n-2)[AvS_w - \min_{v\in\{1,\ldots,n\}}(AvS_v)]}{Max_{v\in\{1,\ldots,n\}}(AvS_v) - \min_{v\in\{1,\ldots,n\}}(AvS_v)} + w + 1. \tag{6}$$

As expected, this penalty function is a balance between the cardinality of the level and the intra-cluster mean distance. The minimum value of the KGS function can be chosen as an autonomous way (as opposite to a user driven way) to prune the dendrogram. AClAP also provides a detailed description of all local minima occurring before the global minimum is reached; this allows the user to adopt other cutting levels, in the search of more homogeneous clusters.

2.4 Cluster significance

The Chauvenet criterion is often used to determine whether the population of a cluster is statistically significant. According to it, a cluster is significantly populated if its cardinality is more than twice the standard deviation apart from the average population value for that level of clustering. Our rationale for the use of this criterion is to assess whether or not there is evidence that significantly populated clusters deserve particular attention in the conformational analysis and docking contexts.

3 RESULTS AND DISCUSSION

AClAP resulted both in an innovative protocol to autonomously partition conformational datasets, and in a program that accomplishes it in a negligible computational time as compared to that needed to generate the dataset itself. On an Intel PentiumIV (2.4 GHz) processor with 512 MB of RAM, AClAP performed CA over 520 conformations of the drug Fexofenadine in 260 sec. In Figure 3, an example of a typical AClAP report is shown. The program provides the overall number of clusters and how many of them are non singleton. For each cluster, the centroid and the representative conformation, as above defined, are calculated. If a reference conformation, such as a crystallographic pose, is available, AClAP allows a comparison of all the representative poses with it (on an RMSD basis).

3.1 KGS penalty function

With standard options, AClAP uses the global minimum of the penalty function as a cutting criterion of the dendrogram. It is

CLUSTER #	CARD	REPR	CEN.DI	RMSD	CHAU
1	2	219	0.445	5.057	no
2	3	298	0.955	4.782	no
3	45	303	1.763	9.827	yes
4	4	305	2.536	9.922	no
5	32	96	3.538	3.070	no
6	12	97	4.458	3.092	no
7	5	244	4.870	5.134	no
8	11	256	5.678	4.925	no
9	11	274	6.635	4.939	no
10	3	257	6.999	4.632	no

Fig. 3. Excerpt from an AClAP report. Columns refer to: cluster number, cardinality and representative conformation; the distance from this latter to the centroid; the RMSD from a reference conformation and whether the cluster is significantly populated.

interesting to comment about the behavior of this function and about the information it provides. We observe that the function P is the sum of two terms: there is a constant slope term that accounts for the increase of the number of clusters and a term that is proportional to AvS_w. One property that would be of interest for the penalty function is a unique minimum. This property would be guaranteed if the average spreads were monotonically decreasing and concave functions. To this aim, let's consider a single step increment of the average spread, which can be reformulated in the following way:

$$AvS_w - AvS_{w-1} = -\frac{1}{w(w-1)}\sum_{M=1}^{w-2} S_M + \\ -\frac{1}{w-1}\frac{2}{(\chi_A+\chi_B)(\chi_A+\chi_B-1)}\sum_{a=1}^{\chi_A}\sum_{b=1}^{\chi_B} d_{a,b} + \\ 0< \left(\frac{S_A}{w}\right)\left[1-\left(\frac{w}{w-1}\right)\frac{\chi_A(\chi_A-1)}{(\chi_A+\chi_B)(\chi_A+\chi_B-1)}\right] + \\ \left(\frac{S_B}{w}\right)\left[1-\left(\frac{w}{w-1}\right)\frac{\chi_B(\chi_B-1)}{(\chi_A+\chi_B)(\chi_A+\chi_B-1)}\right] \tag{7}$$

$$>1 \qquad <1$$

Here, we called A and B the clusters merged at the current clustering step. One can see that the increment is given by the sum of four terms, the first one is always negative, but supposedly small, and is related to the average spread of the clusters non involved in the current step. The second one, again negative, is the average spread given by the inter-cluster (A and B) conformations. Third and fourth terms have no fixed sign but it can be assumed that most of the times they are positive. Given the way the clustering algorithm works, monotonicity and concavity would be implied by a second term being always prevalent over the last two. In general, this is not true. But we gain an interpretation clue from this: any time we see a definite decreasing behavior of the penalty function; it

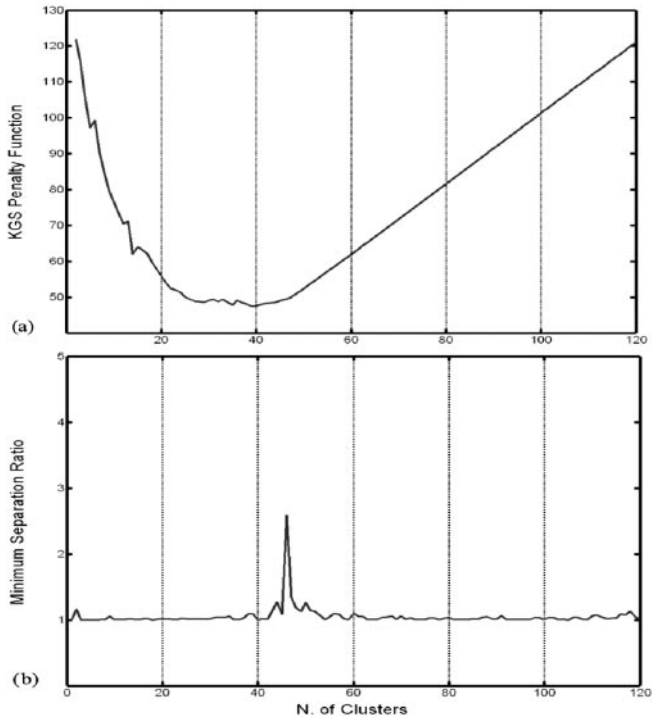


Fig. 4. Cutting rule indicators as implemented in AClAP and X-cluster, they are applied to a member of the ''easy'' drug set: Pimozide (121 conformers). a) The KGS penalty function is plotted vs. the overall cluster number. Standardly, AClAP adopts the minimum of the KGS penalty function as cutting level. b) The MSR value is plotted vs. the overall number of clusters. X-cluster uses the MSR as an indication of different cutting levels. The complete set of clustering results is available as Supplementary Information. These plots show that both algorithms point at a well-defined partitioning.

means that the clustering process has merged two clusters that were well separated. In other words, if one is very much concerned about intra-cluster homogeneity, one has to stop the clustering process at the first pronounced local minimum (which is the rightmost in the plot versus w). Going further on means privileging synthetic representation with respect to intra-cluster homogeneity (see Figures 4a and 6a as typical examples of the KGS behavior).

In what follows, we report the observations we were able to make on the conformational space sampling made both by Metropolis Monte Carlo simulations, and by four docking tools, pointing out how and where they could benefit from the new clustering method.

3.2 Monte Carlo conformational analysis

Metropolis Monte Carlo method (T = 300 K) as implemented in the MacroModel software package was used to perform a conformational analysis on ten marketed drugs. We approximately could split them in conformationally ''easy'' and ''hard'' ones: drugs with up to seven rotatable bonds for which Monte Carlo search provided less than 150 conformers were assigned to the ''easy'' set, whereas the others (with up to 10 rotatable bonds and more than 150 conformers) were defined as ''hard'' compounds.

In the following, we compare the clustering outcomes of AClAP to those of X-cluster (Shenkin and McDonald, 1994),

Table 1. AClAP results for the drug conformations generated via Monte Carlo simulations. For Fentanyl, H^* was 0.53 indicating that CA was not justified. This rule holds more strictly when docking simulations are concerned, whereas drug conformers might still benefit from a CA. Rot. stands for rotatable, NS for NonSingleton and Sign. for significantly populated according to the Chauvenet criterion.

Drug	# confs	Rot. bonds	Max RMSD	H^*	# clusters	NS clusters	Sign. clusters
Conformationally ''easy'' drugs							
Prazosin	24	4	3.24	0.83	5	5	1
Amsacrine	33	5	5.20	0.67	12	11	0
Citalopram	37	4	2.80	0.80	19	14	2
Mizolastine	47	5	4.10	0.65	14	13	0
Fentanyl	49	7	4.26	0.53	12	9	0
Pimozide	121	7	7.30	0.62	39	29	1
Conformationally ''hard'' drugs							
Astemizole	235	8	6.44	0.63	24	24	0
Bepridil	285	9	5.90	0.80	40	35	2
Dofetilide	414	10	10.86	0.77	44	30	3
Fexofenadine	520	9	8.59	0.84	74	58	4
Astemizole	235	8	6.44	0.63	24	24	0

a commonly used clustering procedure implemented in the MacroModel software package. X-cluster is a hierarchical agglomerative clustering method that adopts the single linkage rule. It provides the user with the Minimum Separation Ratio (MSR), which is a function aimed at suggesting a clustering level where all the clusters are well separated. If the MSR is less than 1, the partitioning is expected to be poor. In contrast, an MSR value greater or equal to 2 is an indication of a good partitioning. The final choice of the clustering level is however left to the user.

Preliminarily to our comparison, we adopted the Corrected Rand Index (Hubert and Arabie, 1985) in order to evaluate the similarity of their results. This index is a common measure of the difference between partitionings of the same data set, and it ranges between 0, indicating a strong divergence, and 1, indicating partitioning coincidence.

For the conformationally ''easy'' drugs of the series, Prazosin, Amsacrine, Citalopram, Mizolastine, Fentanyl, and Pimozide, AClAP was able to indicate a functional partitioning, while X-cluster had success in 5 out of 6 cases. AClAP decided for the best clustering level according to the minimum of the KGS penalty function. In Table 1, overall results of AClAP are reported, while Figure 4 shows the AClAP (Figure 4a) and X-cluster (Figure 4b) outcomes applied to the 121 conformers of Pimozide, taken as a representative example for the set of ''easy'' drugs. Figure 4b clearly indicates that, in the reported example, MSR was able to point to a plausible partitioning.

Partitioning obtained by X-cluster applied to conformationally ''easy'' drugs is summarized in Table 2. In particular, for Prazosin, Citalopram, and Pimozide, the MSR values pointed univocally to a cutting level for the hierarchical tree. The partitionings strongly agree with those obtained by AClAP, the Corrected Rand Index values being 0.74, 0.79, and 0.79, for the three molecules, respectively (see the last column of Table 2). Conversely, for Fentanyl, the MSR value provided no clear indication of a cutting level. The H^* value for Fentanyl provided by AClAP was 0.53 (see

Table 2. X-cluster results for 5 ''easy'' drugs, whose conformations were generated via Monte Carlo simulations. X-cluster did not provide any significant cutting point for Fentanyl.

Drug	# conformers	MSR	# clusters	NS Cluster	Corrected Rand Index
Prazosin	24	1.90	4	4	0.74
Amsacrine	33	4.42	2	2	0.03
Citalopram	37	19.40	21	16	0.79
Mizolastine 1	47	2.00	2	2	0.15
Mizolastine 2	47	1.93	18	14	0.90
Pimozide	121	2.58	47	33	0.79

Table 1), suggesting that conformations did not display a natural tendency to aggregate into groups. It should be mentioned that, when H^* is less than 0.6, unlike structure-based drug design, ligand-based drug design might still benefit from CA applied to drug conformers. Consistently, AClAP applied on Fentanyl provided a quite good partitioning, as reported in Table 1. In the case of another ''easy'' drug, Amsacrine, a significant MSR value led to a partition with only two clusters. Conversely, the partition provided by AClAP afforded 12 clusters. The Corrected Rand Index was as low as 0.03, indicating that the two partitionings were markedly different (see Figures 5a and 5b). As it can be seen in Figure 5b, the internal homogeneity of the partitioning provided by X-cluster was rather poor. One possible reason could be the chaining effect induced by the single linkage rule. Finally, in the analysis of Mizolastine, two clustering levels worthy to be selected were identified, showing MSR values of 2 and 1.93, respectively (Mizolastine 1 and Mizolastine 2 in Table 2). The clustering of Mizolastine 1 (MSR = 2) corresponded to a partition with only two clusters, lacking internal homogeneity and displaying an evident chaining effect (data not shown). The second partitioning (Mizolastine 2, 18 clusters, 14 nonsingletons) provided more homogenous clusters and a strong agreement with the partition obtained by AClAP (Corrected Rand Index = 0.90).

When processing the conformationally ''hard'' drug set (composed by Astemizole, Bepridil, Dofetilide, and Fexofenadine), whose conformers were generated via Metropolis Monte Carlo simulations, X-cluster did not provide any clue about the cutting level for the conformations, demonstrating that a protocol based on the single linkage rule in combination with MSR fails when dealing with conformationally complex molecules. In Figure 6, as an example, the 520 conformers of Fexofenadine treated with AClAP (Figure 6a) and X-cluster (Figure 6b) are shown. As reported in Table 1, in these cases H^* test showed a natural grouping tendency, and AClAP, a protocol based on the average linkage rule in combination with the KGS penalty function was actually able to univocally provide a good partitioning for all the drug conformers (see Figure 6a and Table 1). We can conclude that, for the drugs here investigated, AClAP definitely outperformed X-cluster.

3.3 Docking simulations

We studied the conformational sampling done by four among the most widespread docking programs, namely, Dock, AutoDock,

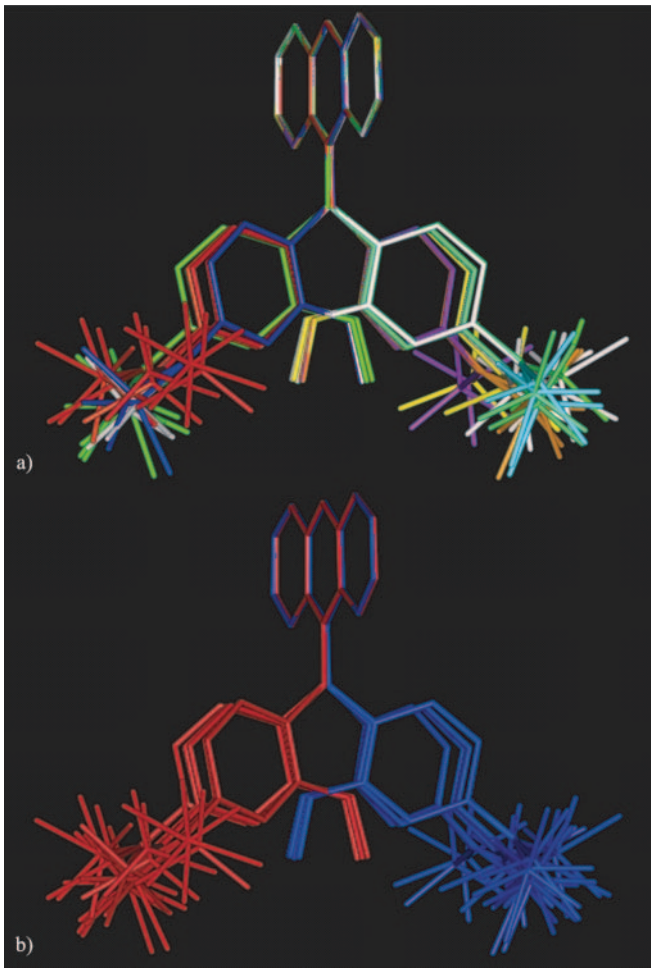


Fig. 5. The partitioning of 33 conformers of Amsacrine. **a**) The partitioning provided by AClAP. A protocol based on the average linkage rule and the KGS cutting function generated 12 groups bearing high intra-cluster homogeneity. **b**) The partitioning provided by X-cluster. A protocol based on the single linkage rule and a user-dependent cutting function generated 2 groups bearing scarce intra-cluster homogeneity.

GOLD, and FlexX, together with the action of our clustering protocol over their output. We ran the programs over a set of 16 crystallographic complexes belonging to the following protein families: kinases, hormone receptors, and proteases (both serine and aspartic proteases). As a figure of merit, we took the RMSD of the generated poses from the crystallographic one. For a detailed description of docking simulations and comparative analysis the reader is referred to the work of Bottegoni *et al*. (Bottegoni, *et al*., 2006). In what follows, we summarize some conclusions we drew from that experience. The present comments encompass only 15 cases, since one of the original ones (Propidium co-crystallized with AChE, PDB code 1N5R) has been demonstrated to bind to the surface of its biological counterpart in at least two different modes (Bourne, *et al*., 2003; Cavalli, *et al*., 2004).

About conformational sampling, and having defined a ''good'' pose as the one which is less than 2.5 Å far away (in terms of RMSD

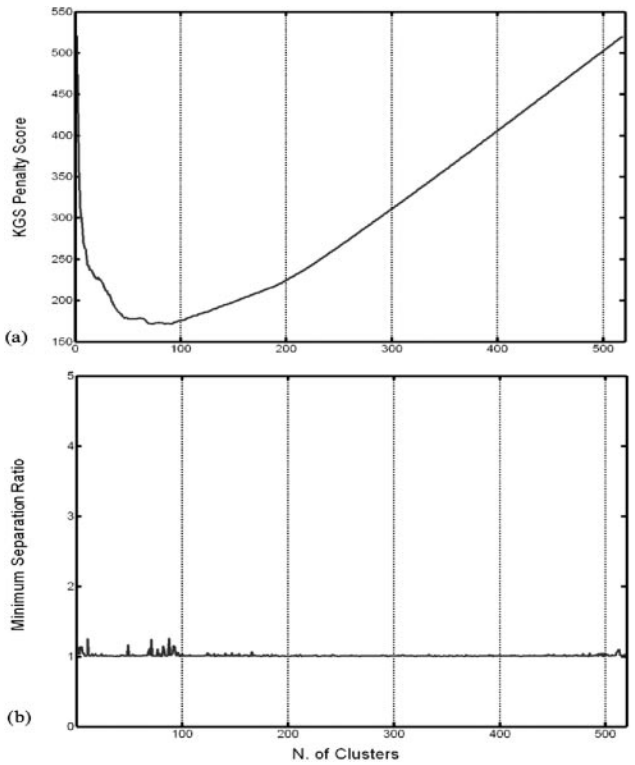


Fig. 6. The cutting rule as implemented in AClAP and X-cluster. As an example of complex, ''hard'', compound, the CA was performed over 520 conformers of Fexofenadine. **a**) The KGS penalty function is plotted vs. the overall number of clusters. Standardly, AClAP adopts the minimum of the KGS penalty function as cutting level. **b**) The MSR value is plotted vs. the overall number of clusters. X-cluster uses the MSR as an indication of possible cutting levels. These plots show that only the KGS score was able to point at a well-defined partitioning. The complete set of clustering results is available as Supplementary Information.

between nonhydrogen atoms) from the crystallographic one, we can comment as follows:

- at least one among the docking tools was always able to generate a pose sufficiently close to the experimental one, being of 89% the average success rate among docking programs; in particular, AutoDock and GOLD were always able to provide at least one ''good'' pose, whereas Dock and FlexX had a success rate of 80% and 73%, respectively;
- however, in terms of conformational sampling, no docking tool significantly outperformed the others, with a chi-square value of 0.67, corresponding to a 88% of probability that fluctuations in the results are random.

Comments about clustering features, shown in more detail in Table 3, follow:

- the best pose was found within a singleton cluster with a very low frequency, ranging from 0.2% to 0.7%;
- when a single docking tool was used, the assertion that ''good'' poses are to be found only, or mainly, in the most populated clusters did not find any clear evidence;
- when a ''holistic'' approach was adopted, i.e., the clustering was performed over the poses generated by all of the docking tools, the probability of finding at least one ''good'' pose among the representative conformations of the most populated clusters, whose number was always between 1 and 3, reached roughly 93%;
- in the holistic approach, as compared with the single tool approach, the presence of ''good'' poses decreases in scarcely populated clusters in favor of very highly populated ones.

A comment is due about the performance of AClAP in the so-called holistic approach. No scoring process was used to support the provided results. Nevertheless, their performance can be compared to that of the widely used consensus scoring method, which well overcomes the main limitation of scoring functions. Indeed, also in our investigation, the scoring functions sometimes failed to rank correctly the best poses: roughly in the 50% of the cases. We found of particular interest the data shown in the last two columns of Table 3: they indicate that, at least for the molecules we examined, there is a high chance to find a ''good'' pose among the representative conformations of the most populated clusters. According to our arrangement procedure in bins, and similarly to what obtained with the Chauvenet criterion, those conformations are usually less than two. This procedure seems to point at a few, but still very promising, candidates that can be successively examined with more accurate tools, providing a really remarkable dimensionality reduction.

4 CONCLUSIONS

In this paper, we have described a new clustering protocol as well as its implementation in a MATLAB program. The new software, named AClAP, turned out to be well suited to cluster both conformations generated via Metropolis Monte Carlo simulations of drugs, and poses obtained by reiterated docking runs. In a consistent fashion, AClAP prompts the user to assess the clusterability of a conformational dataset by means of what we named the H^* test. The subsequent step is a hierarchical agglomerative cluster analysis based on the average linkage rule. The choice of this rule with respect to others was already discussed elsewhere (Bottegoni, *et al*., 2006), and here reinforced. Once the hierarchical tree is built, an autonomous method to prune it is needed to define the best clustering level. Here, we have shown that the KGS penalty function is an unbiased approach very well suited to achieve that goal. AClAP outperforms standard CA-based protocols as they are implemented in the most commonly used docking programs. In this context, the AClAP method manages to greatly reduce conformational space dimensionality, proving to be fruitful, for instance, for the successive application of computationally intensive energy estimation techniques to be applied to cluster representatives. On top of it, in what we called the holistic approach, AClAP allowed us to identify some one among the closest poses to the experimental one, and placed it within a statistically significant cluster with a very promising hit rate. Finally, when applied to the output of Metropolis Monte Carlo searches, AClAP proved to be more robust than the long-time exploited and commonly used X-cluster routine. Encouraged by the present results, we propose AClAP as a new and user-friendly tool to help molecular modelers facing issues related to both ligand- and target-based drug design. Our efforts are currently devoted to extend the appli-

Table 3. Distribution of ''good'' poses with respect to relative cluster cardinality. For each docked molecule, at the clustering level chosen by AClAP, the clusters are classified in five bins (A to E) according to their cardinality, in ascending order. Then, for each molecule, the relative frequency of the ''good'' poses, i.e., those with an RMSD < 2.5 Å from the crystallographic one, is calculated. The derived frequencies, at fixed docking program, are then averaged over the different molecules. In the last row, the outcome of the holistic approach is reported. In parentheses, the bin population with respect to the total cluster number is shown. As one can see, the ''good'' poses tend to be distributed among very highly and very scarcely populated clusters, with a prevalence of the formers. The holistic approach seems to make this prevalence maximally marked.

	Class A (least populated clusters) %	Class B %	Class C %	Class D %	Class E (most populated clusters) %	Good poses in singleton clusters %	Frequency of at least one ''good'' pose in a signif. populated cluster %.	Frequency that at least one representative pose of a cluster in E bin is ''good'' %	Average number of clusters in E bin
AutoDock	12.6 (86.6)	9.0 (5.7)	6.4 (1.6)	7.6 (1.5)	64.4 (4.6)	0.3	80.0	80.0	1.1
FlexX *	33.0 (72.0)	23.8 (14.1)	5.5 (6.2)	0.8 (2.2)	36.9 (5.5)	0.2	33.3	40.0	1.3
Dock *	31.2 (83.4)	10.8 (4.4)	5.9 (1.1)	2.2 (1.0)	49.9 (10.1)	0.7	53.3	50.0	1.1
GOLD	14.9 (76.1)	7.3 (7.1)	18.3 (5.0)	1.8 (0.2)	57.7 (11.6)	0.5	40.0	78.6	1.2
Average	**22.9 (79.5)**	**12.7 (7.8)**	**9.0 (3.5)**	**3.1 (1.2)**	**52.2 (8.0)**	**0.4**	**51.7**	**62.1**	**1.2**
Holistic	**14.3 (94.3)**	**6.8 (2.3)**	**11.1 (1.2)**	**2.3 (0.4)**	**65.5 (1.8)**	**0.2**	**100.0**	**93.3**	**1.1**

*In one case, these programs weren't able to find any pose closer than 2.5 Å to the crystallographic one.

cability of this approach to rationalize the outcomes of protein-protein docking.

ACKNOWLEDGEMENTS

We thank M. Masetti for technical assistance. Miur-Cofin2004 and FIRB projects (''Sviluppo di metodologie innovative per l'identificazione e la sintesi di nuove molecole a scopo terapeutico: applicazioni nel campo della malattia di Alzheimer'' and ''Laboratorio Nazionale sulle Nanotecnologie per Genomica e Postgenomica (NG-Lab)'') are gratefully acknowledged for the financial support.

REFERENCES

Bottegoni,G., Cavalli,A. and Recanatini,M. (2006) A comparative study on the application of hierarchical-agglomerative clustering approaches to organize outputs of reiterated docking runs. *J. Chem. Inf. Mod.*, **46**, 852–862.

Bourne,Y., Taylor,P., Radic,Z. and Marchot,P. (2003) Structural insights into ligand interactions at the acetylcholinesterase peripheral anionic site. *Embo J.*, **22**, 1–12.

Cavalli,A., Bottegoni,G., Raco,C., De Vivo,M. and Recanatini,M. (2004) A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase. *J. Med. Chem.*, **47**, 3991–3999.

Chang,G., Guida,W.C. and Still,W.C. (1989) An internal coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.*, **111**, 4379–4386.

Chema,D., Eren,D., Yayon,A., Goldblum,A. and Zaliani,A. (2004) Identifying the binding mode of a molecular scaffold. *J. Comput. Aided Mol. Des.*, **18**, 23–40.

Everitt,B.S., Landau,S. and Leese,M. (2001) Cluster Analysis. Arnold, a member of the Hodder Headline Group,London.

Ewing,T.J., Makino,S., Skillman,A.G. and Kuntz,I.D. (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, **15**, 411–428.

Hopkins,B. (1954) A new method for determining the type of distribution of plant individuals. *Ann. Bot.-London*, **18**, 213–227.

Hubert,L.J. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Jones,G., Willett,P., Glen,R.C., Leach,A.R. and Taylor,R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, **267**, 727–748.

Kelley,L.A., Gardner,S.P. and Sutcliffe,M.J. (1997) An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. *Protein Eng.*, **10**, 737–741.

Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Mohamadi,F., Richards,N.G.J., Guida,W.C., Liskamp,R.M.J., Lipton,M.A., Caulfield,C.E., Chang,G., Hendrickson,T.F. and Still,W.C. (1990) MacroModel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.*, **1**, 440–467.

Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. and Olson,A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.

Rarey,M., Kramer,B., Lengauer,T. and Klebe,G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.

Shenkin,P.S. and McDonald,D.Q. (1994) Cluster analysis of molecular conformations. *J. Comput. Chem.*, **15**, 899–916.

Taylor,R.D., Jewsbury,P.J. and Essex,J.W. (2002) A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.*, **16**, 151–166.

Ward,J.H.J. and Hook,M.E. (1963) Application of a hierarchical grouping procedure to problem of grouping profiles. *Educ. Psychol. Meas.*, **23**, 69–92.