A Comparative Study on the Application of Hierarchical−Agglomerative Clustering Approaches to Organize Outputs of Reiterated Docking Runs

Giovanni Bottegoni, Andrea Cavalli,* and Maurizio Recanatini

Department of Pharmaceutical Sciences, University of Bologna, Via Belmeloro, 6-I-40126 Bologna, Italy

Received April 21, 2005

Reiterated runs of standard docking protocols usually provide a collection of possible binding modes rather than pinpoint a single solution. Usually, this ensemble is then ranked by means of an energy-based scoring function. However, since many degrees of approximation have to be introduced in the computation of the binding free energy, scoring functions cannot always rank the experimental pose among the top scorers. Cluster analysis might help to overcome this limit, provided that data clusterability has been earlier assessed. In this paper, first, we present a modified version of a test earlier developed by Hopkins to assess whether or not docking outputs show the natural tendency to be grouped in clusters. Then, we report the results of a comparative study on the application of different hierarchical−agglomerative cluster rules to partition docking outputs. The rule that was able to best manage the observed data was finally applied to the whole ensemble of poses collected from several docking tools. The combination of the average linkage rule with the cutting function developed by Sutcliffe and co-workers turned out to be an approach that meets all of the criteria required for a robust clustering protocol. Furthermore, a consensus clustering allowed us to identify the pose closest to the experimental one within a statistically significant cluster, whose number was always of few units.

INTRODUCTION

In the structural genomic era, three-dimensional data on targets of pharmacological relevance are rapidly increasing, thanks to high-throughput protein structure determination, homology modeling, and de novo structural prediction.[1] Docking and virtual screening are computational methods able to dramatically speed up the hit identification process and are playing a central role in modern structure-based drug design.[2−4]

A standard docking procedure consists of two steps, posing and scoring. In the first step, a searching algorithm predicts various ligand conformations and orientations within a target active site. In the second one, a scoring function assesses the ligand−protein interactions' tightness through the estimation of the binding free energy of each docked pose. Virtual screening can be considered as the application of a docking scheme over a database of compounds. In virtual screening simulations, the scoring function ranks as top scorers those ligands forming the most stable complexes with the biological counterpart.

A large amount of published data,[5−8] mainly based on comparative studies on several docking programs and scoring functions, allows one to lay down general guidelines. First, searching algorithms are usually able to reproduce experimental poses with a root-mean-square deviation (RMSD) less than 2.5 Å, despite the fact that a deep exploration of the conformational space is not a trivial task to accomplish in a reasonable amount of computer time.[9] Second, scoring functions do not always succeed in including the experimental pose among the most favorable ones.[10] Actually, currently available scoring functions, either force-field-, experimental-, or knowledge-based, introduce relevant degrees of approximation, particularly when dealing with entropic and solvation contributions.[3] On the other hand, sampling-based methods, such as free energy perturbation, umbrella sampling, and so forth, would be much more accurate in predicting binding free energy, being however computationally too expensive to be routinely applied in docking studies.[11] In this context, it is worth mentioning the docking protocol very recently proposed by Gervasio et al.[12] This metadynamics-based approach was faster than common sampling methods, being however able to predict experimental poses and binding free energies, mimicking a ligand exiting or entering a target active site.

In light of the above considerations, it is not surprising that standard docking protocols often provide a collection of possible binding modes rather than pinpoint a single solution. In this context, it should be highlighted that docking suites relying upon a genetic algorithm provide by definition an ensemble of results. Conversely, deterministic algorithms often miss the ability to converge to a global minimum starting from different docking poses.[13] Eventually, scoring functions should rank the poses in terms of the free energy of binding. However, if the reliability of a scoring function is questioned, any of the provided poses might be, in principle, accepted or discarded and, therefore, need to be further investigated. Since the poses ensemble is an actual collection of observed data, it may be, for instance, organized by means of a statistical approach.

Cluster analysis is a powerful tool to organize observed data into meaningful subsets (clusters). Clustering approaches are becoming consolidated methods to analyze docking results and to assess whether the poses can be partitioned in

* Corresponding author phone: +39 051 2099735; fax: +39 051 2099734; e-mail: andrea.cavalli@unibo.it

10.1021/ci050141q CCC: $33.50
Published on Web 02/09/2006

terms of clusters.[14,15] A critical step in the application of a cluster analysis is to choose a protocol able to provide a functional partition. That is, clusters should be internally homogeneous, and the partition should be performed without a priori knowledge of the most significant solution. Moreover, clusters should reflect the real trend of the explored conformational space, while, skipping irrelevant differences, making it possible to identify those conformations worthy of further study and consequently rejecting the outliers. Finally, the unsupervised partition of the data should be rigid, not allowing overlapping clusters, and robust in terms of reproducibility. However, the possibility exists that a docking protocol provides an ensemble of poses bearing slight conformational differences in the binding mode. This may be due to the presence of a low number of rotatable bonds on the ligand, a narrow active site, and in general a situation when the ligand−protein complex is located in a very deep well of the free energy surface. In this case, docking poses simply correspond to a single homogeneous collection, and the application of a cluster analysis leads to artifacts. The statistical approach would force a pattern to the docking results rather than discover it, because a cluster analysis partitions data whether or not they display a natural clustering pattern. Therefore, a preliminary assessment of the data partitioning (data clusterability) should always be part of an accurate clustering protocol. Such an assessment has to be able to a priori identify the presence of natural groups without necessarily partitioning them.

In this paper, we present a comparison of three different hierarchical−agglomerative clustering methods to partition the outputs of four common docking suites. The clusterability of the docking poses was preliminarily assessed by means of a modified version of the Hopkins test.[16] The clustering level (i.e., the cutting level) was determined by means of the penalty score function developed by Sutcliffe and coworkers.[17] The rule that is able to best manage the observed data is finally applied to the whole ensemble of poses collected from several docking tools. Such an integrated methodology ended up matching all of the criteria required for a proper clustering procedure and greatly improving the reliability of molecular docking outputs.

METHODS

Cocrystal Complexes. Docking and clustering simulations were carried out on a set of 16 protein−ligand cocrystals, retrieved from the Protein Data Bank (PDB). The complexes were collected according to the following criteria: (i) the ligands do not form covalent bonds with the proteins, (ii) the ligands have a low number of rotatable bonds, and (iii) the crystallographic resolution was less than 2.80 Å. We chose cocrystals belonging to three different protein families of pharmaceutical interest (kinase, hormone receptor, and protease), thus ensuring enough diversity for the purposes of the present work. The complete list of the crystallographic complexes employed in the study is reported in the Supporting Information.

Once retrieved from the PDB and extracted from the crystallographic complexes, ligands were properly modified by means of the SYBYL 7.1 molecular modeling suite (Tripos Inc., St. Louis, MO). The protonation state of the ligands was defined according to the actual pK_a of the molecules. Correct atom types and bonds order were defined, hydrogen atoms were added, and charges were loaded by means of the AM1 semiempirical Hamiltonian,[18] as implemented in the SYBYL graphical interface to MOPAC. During the input phase, all docking programs were able to automatically detect both torsional degrees of freedom and rigid fragments (where requested).

Docking Protocols. In this study, cluster analysis was carried out on the outputs of four docking programs: AutoDock 3.0.5, GOLD 2.1.2, Dock 4.0.1, and FlexX 1.13. GOLD and FlexX were available to us under demo licenses. Besides being among the most commonly employed docking programs,[13] AutoDock[19] and GOLD[20] are genetic-based algorithms, whereas Dock[21] and FlexX[22] are fragment-based deterministic software.

AutoDock. Targets were described with a united-atom model, and the Kollmann charges were loaded using the AutoDockTools graphical interface. Grid maps were centered on the ligand and their dimensions set to $60 \times 60 \times 60$ points with a spacing of 0.375 Å. Genetic algorithm local search parameters were set as follows: the initial population was made by 50 individuals, the maximum number of energy evaluation was 2.5×10^5, and the maximum number of generation was 2.7×10^4. Other parameters were kept as provided by default. For each complex, 100 docking runs were carried out.

GOLD. Ligand geometric centroid coordinates were selected as coordinates of the active-site origin. The active-site radius was set equal to 10 Å. As suggested by the GOLD authors,[20] genetic algorithm default parameters were set: the population size was 100, the selection pressure was 1.1, the number of operations was 10^5, the number of islands was 5, the niche size was 2, migrate was 10, mutate was 95, and crossover was 95.

Dock. Targets were described with a united-atom model, and the Kollmann charges were loaded using the Biopolymer module of SYBYL 7.1. All residues bearing at least one heavy atom within 6.5 Å from the ligand composed the receptor active site. The active-site surface was described using a spherical probe with a radius of 1.4 Å and a surface density of 8.0 dots/Å^2. The docking spheres were generated with the SPHGEN program, while the Grid program calculated the energy score. The grid spacing and the energy cutoff were set to 0.4 and 10 Å, respectively. A total of 100 flexible docking runs were carried out changing the random number generator seed and storing one pose in the output. Docking parameters were set as follows: 50 configurations per cycle, 500 maximum configurations, minimization maximum cycles set to 1, maximum iteration set to 100, dielectric factor set to 4, and energy cutoff set to 10 Å.

FlexX. Active-site coordinates were defined using the ligand X-ray structure. The cutoff value was set to 6.5 Å. Base fragments were automatically selected by means of an internal scoring scheme. The maximum number of base fragments was set equal to 5. Base fragments were placed in the active site using a standard algorithm based on triangle hashing techniques. If the base fragments had fewer than three interaction centers, a second algorithm, based on the matching of line segments, was used. Other parameters were kept as provided by default. For each complex, 100 docking runs were carried out.

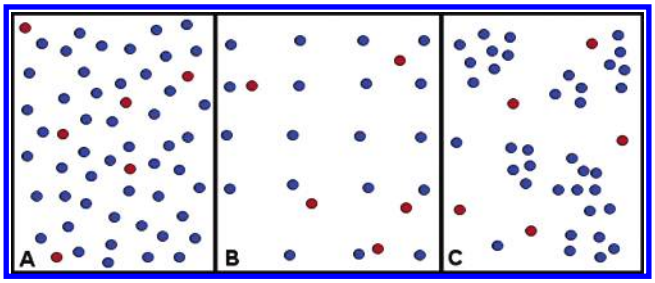

Figure 1. H* test for the clusterability assessment. Real and virtual points are blue and red, respectively. (A) Both virtual and real points are homogeneously distributed in the space. (B) Virtual points are homogeneously distributed in the space, while the real ones are regularly spaced. (C) Virtual points are homogeneously distributed in the space, while the real ones show a natural tendency to cluster.

Clusterability Assessment. To preliminarily assess the clusterability of docking outputs, a simple test (H* test) based on the work of Hopkins[16] was developed. The distribution of a data set can be (i) uniformly scattered, (ii) regularly spaced, or (iii) naturally grouped. Indeed, only in the last case should a cluster analysis should be performed.

The H* test was composed of the following steps: (i) construction of a bidimensional matrix, (ii) principal component analysis (PCA) on the variables of the matrix elements, (iii) plotting of the principal components in the components space, (iv) random choice of five components (real points) among the plotted ones, (v) generation of random points (virtual points) within the components space, and (vi) calculation of the H* value.

In detail, each pose is considered as an observation in a *d*-dimensional space and stored in an *n*-by-*d* matrix **M**, where *n* is the number of the sampled poses (100) and *d* is the number of variables. In this study, for each molecule, the variables were defined as both the Cartesian coordinates of three atoms that accounted for the rotational and translational degrees of freedom and the dihedral angle values of all rotatable bonds that accounted for the internal conformational degrees of freedom. For each of the *d* columns, the mean and standard deviation was calculated to normalize the matrix elements *x* as follows:

$$x_{\mathrm{NORM}-ij} = \frac{x_{ij} - (\mathrm{mean}_j)}{\mathrm{std}_j} \qquad (1)$$

In eq 1, $1 \le i \le 100$, $1 \le j \le d$, and mean_j and std_j are the mean value and the standard deviation of the *j*th column, respectively. Equation 1 assured homogeneity of the dimension of the variables (i.e., Cartesian coordinates and torsional angles of all 100 docking poses).

A PCA was then performed on **M**. A variance−covariance matrix **Z** was obtained:

$$\mathbf{Z} = \mathbf{M}^{\mathrm{T}}\mathbf{M} \qquad (2)$$

In eq 2, $\mathbf{M}^{\mathrm{T}}$ is the transpose of **M**. The eigenvectors of **Z** are the coefficients of the principal components and can be obtained, along with their associated eigenvalues, by matrix diagonalization. In eq 2, **Z** is a square symmetric matrix, and its eigenvectors are orthogonal. In the H* test, the principal components corresponded to the *d* dimension(s) of the multidimensional space. In Figure 1, an example with only two components, corresponding to the *x* and *y* axes of a plane, is reported.

In this example, the components describing the variables of the docking poses were represented as points in a bidimensional space (the real points, blue in Figure 1). The H* test then proceeded by randomly selecting five real points. The Euclidean distance (*D*) between each selected real point and its nearest neighbor of the whole ensemble was calculated. Then, five virtual points (red in Figure 1) were generated. The coordinates of the virtual points were sampled from a normal distribution with the mean set to zero and the standard deviation set to the standard deviation of the principal components. The Euclidean distance (*V*) between each virtual point and its nearest neighbor among the real points was calculated. The H* value was finally obtained solving the following equation:

$$\mathrm{H}^* = \frac{\sum_{i=1}^{5} V_i}{\sum_{i=1}^{5} V_i + \sum_{i=1}^{5} D_i} \qquad (3)$$

Solving eq 3, one may sum up three main cases.

(*i*) $V \approx D \Rightarrow 0.5 \le \mathrm{H}^* \le 0.6$, the real points are homogeneously distributed (Figure 1A)

(*ii*) $V \ll D \Rightarrow \mathrm{H}^* \to 0$, the real points are regularly spaced (Figure 1B)

(*iii*) $V \gg D \Rightarrow \mathrm{H}^* \to 1$, the real points show a natural tendency to clustering (Figure 1C)

A cluster analysis should be carried out only when $\mathrm{H}^* \to 1$. In the analysis of docking outputs, the $\mathrm{H}^* \to 0$ case is not expected to occur. This means that docking poses are assumed to never be placed at the protein active site with a regular and repeated displacement.

Cluster Analysis. A hierarchical−agglomerative script was purposely written and developed in our laboratory, to rationalize docking outputs. "Hierarchical" means that clusters at a high clustering level are unions of clusters at lower levels, while "agglomerative" means that clusters never break apart during the formation process.

First, a dissimilarity matrix (a one-mode square symmetric matrix) is calculated, where its elements are the RMSDs between each pair of the 100 poses provided by every docking protocol. The clustering algorithm then proceeds with a series of partitions starting with *N* clusters, each containing a single docking pose (where *N* is the total number of poses, in this study, 100), and ending with one cluster containing them all. Three alternative linkage rules were implemented: single linkage, average linkage, and the Ward method, which share similar basic operations. In particular, at each stage, the linkage rule clusters the closest poses. The three rules differ in the way similarity is determined. In detail, when the single linkage,[23] also known as the nearest-neighbor distance, is used, the similarity between groups is calculated using the closest pair of poses, while with the average linkage[23] the mean distance between all pairs of poses is measured. The Ward method uses an analysis of variance approach to evaluate distances.[24]

The global hierarchy comprises all of the possible clustering solutions in a nested sequence, usually referred to as

dendrogram, spanning from N to 1 cluster. Arbitrarily choosing a functional partition (the cutting level), namely, the number of final clusters, may represent a serious drawback for the entire protocol. In this study, the partition was selected by applying the Kelley−Gardner−Sutcliffe (KGS) penalty function to the dendrogram.[17] The method is thoroughly described in the paper by Kelley et al.[17] and here briefly summarized. An average spread value is calculated for each clustering level of the dendrogram. When all average spread values are collected, they are normalized to lie between 1 and $N - 1$. For each ith clustering level, a penalty (P) value is calculated as

$$P_i = \text{AverageSpreadNormalized}_i + n_i \qquad (4)$$

where n_i is the number of clusters at the ith stage.

The minimum value of the KGS function determined, in a fast, elegant, and strictly mathematical manner, the partition that best balances a low number of clusters and the intracluster homogeneity.

RESULTS AND DISCUSSION

The simulations were carried out on a set of 16 cocrystals using four different docking programs and storing 100 poses for each protocol (overall, 6400 poses were analyzed). Once the ensembles of poses were generated, the clusterability was assessed by means of a modified version of the Hopkins test[16] (H* test, see Methods). Then, several cluster analyses were performed employing combinations of different algorithms.

It was assumed that a docking program was able to correctly reproduce the binding mode of a ligand−target complex, if at least one docked pose laid within a RMSD of 2.5 Å from the crystallographic solution. A similar criterion was previously adopted in a comparative docking study.[10] A correct prediction was achieved for 55 out of 64 cases. A reliable binding mode was not obtained for 1N5R docked using GOLD; for 1DI8, 1IEP, 1S9P, and 1N5R docked using Dock; and for 1A28, 1N5R, 1UT6, and 1W51 docked using FlexX (see the Supporting Information for further details on the complexes). All 16 complexes were reliably reproduced by means of the AutoDock program. A further assessment of the different algorithms' docking performance,[6] as well as a detailed investigation of the reasons for failure,[7] were beyond the scope of the present paper. With the only exception being the simulation of 1AGW with FlexX, when several poses were identified, the best energy scorer was never the closest to the experimental one in terms of RMSD. Indeed, this was not surprising, because the difficulty of a docking algorithm to reproduce the crystallographic pose on the basis of scoring functions is widely reported in the literature.[3,9,10]

Clusterability Assessment. The H* test was applied to each data set as a preliminary assessment of the data clusterability. In Table 1, the H* values along with the maximum RMSD (MaxRMSD) calculated among all of the docked poses are reported.

Only in 7 (bold in Table 1) out of 64 cases was the H* value between 0.5 and 0.6. This means that, for as many as 57 out of 64 docking outputs, a postprocessing approach to organize the results would have been greatly recommended. As an example, the dockings of H717 at the CDK2 binding site (PDB code 1G5S) are reported in Figure 2.

Table 1. Clusterability Assessment by Means of the H* Test[a]

	AutoDock		GOLD		Dock		FlexX	
	Max-RMSD	H*	Max-RMSD	H*	Max-RMSD	H*	Max-RMSD	H*
1URW	13.00	0.71	11.61	0.76	4.26	0.81	6.85	0.81
1DI8	14.37	0.81	8.41	0.78	1.59	0.76	15.65	0.82
1G5S	1.10	**0.55**	2.70	**0.56**	1.98	0.84	11.25	0.86
1E1X	13.17	0.84	2.67	**0.58**	9.79	0.83	20.94	0.87
1AGW	12.65	0.72	11.42	0.91	11.06	0.84	20.10	0.71
1Q5K	12.31	0.62	9.43	0.75	2.39	0.67	11.49	0.90
1IEP	14.47	0.66	3.20	0.64	32.61	0.79	25.69	0.90
3ERT	11.96	0.82	2.51	**0.54**	3.00	0.71	2.13	0.77
1S9P	7.25	0.81	7.26	0.87	10.94	0.62	16.10	0.79
1A28	0.56	0.84	0.34	**0.51**	0.67	**0.58**	23.28	0.83
1EVE	11.57	0.81	10.65	0.71	12.23	0.80	11.57	0.76
1N5R	12.26	0.67	7.48	0.76	10.98	0.85	12.94	0.83
1UT6	12.36	0.78	5.22	0.72	7.97	0.80	18.67	0.68
1UV6	23.4	0.79	5.41	0.64	4.85	0.84	14.72	0.88
1W51	15.81	0.65	11.85	0.62	12.58	0.78	7.73	0.71
1NNB	14.95	0.73	1.17	**0.56**	10.94	0.82	8.27	0.76

[a] The H* values along with the maximum RMSD (MaxRMSD) calculated among all docked poses are reported. In bold are the seven cases in which the H* value was between 0.5 and 0.6.

In statistical terms, AutoDock, GOLD, Dock, and FlexX docked the ligand in four different ways, allowing us to explore two ($0.5 < H^* < 0.6$ and $H^* \rightarrow 1$) of the three possible scenarios accounted for by the H* test.

When H* was between 0.5 and 0.6 (Figures 2A−D), a cluster analysis had not to be performed, and two possible subscenarios were identified. First, if the maxRMSD was less than 2 Å, the poses corresponded to a single binding mode, as shown by the docking of H717 at the active site of CDK2 by means of AutoDock (see Figure 2A and Table 1). In Figure 2B, the PCA on the variables of the AutoDock poses is shown. It turned out that the randomly generated virtual points were homogeneously located among the real points plotted against the principal components (for details, see the H* test in the Methods). For the sake of clarity, three components (i.e., a three-dimensional space), which always accounted for more than 80% of the overall variance, are reported. However, the H* test was carried out with up to five components, which always accounted for more than 95% of the variance, provided very similar H* values (data not shown). Second, if the maxRMSD was more than 2 Å, the poses represented a single binding mode, despite a regular rotation of certain dihedral angles of the ligands. This was clearly shown when docking H717 by means of GOLD (see Figure 2C and Table 1). In Figure 2D, the PCA on the variables of the GOLD poses is shown. It can be seen that both virtual and real points were uniformly distributed in the space, again pointing out that a cluster analysis had not to be performed.

In contrast, when H* was greater than 0.6 (i.e., $H^* \rightarrow 1$, Figures 2E−H), the data set exhibited a predisposition to cluster into natural groups. Again, two subscenarios could be identified. First, if the maxRMSD was less than 2 Å, most of the poses corresponded to a single binding mode, while outliers were kept separated. This was the case of the docking of H717 by means of Dock (see Figure 2E and Table 1). When the PCA results (Figure 1F) are plotted, it can be seen that a wide empty space separated a highly populated region from scattered points. Since virtual points were by definition homogeneously distributed, they were prevalently placed in

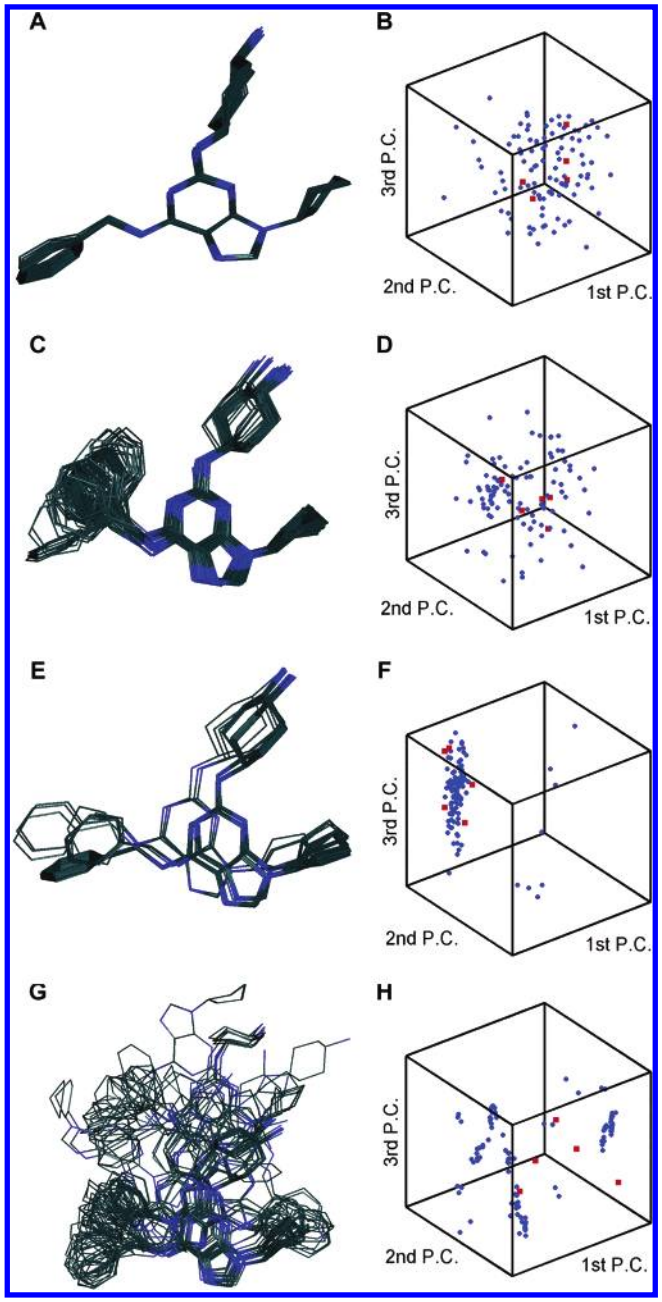

Figure 2. Case study for the H* test: the docking of H717 at the active site of CDK2. Docking with (A) AutoDock, (C) GOLD, (E) Dock, and (G) FlexX. H* test on the (B) AutoDock, (D) GOLD, (F) Dock, and (H) FlexX outputs. Three components accounting for more than 80% of the overall variance are reported. Virtual points and real points are represented as red squares and blue circles, respectively.

the empty space, providing a high H* value. Second, when the maxRMSD was more than 2 Å, several separated groups of poses were identified. This scenario was observed when docking H717 by means of FlexX (see Figure 2G and Table 1). In Figure 2H, the PCA on the variables of the FlexX poses is shown. Again, the virtual points were placed in empty regions among the clusters, providing an H* value close to 1.

We conclude that, when H* is between 0.5 and 0.6, a cluster analysis has not to be performed. In the example, this was the case of the docking of H717 by means of both AutoDock (Figures 2A, B) and GOLD (Figures 2C, D). In contrast, when H* is more than 0.6, a cluster analysis should

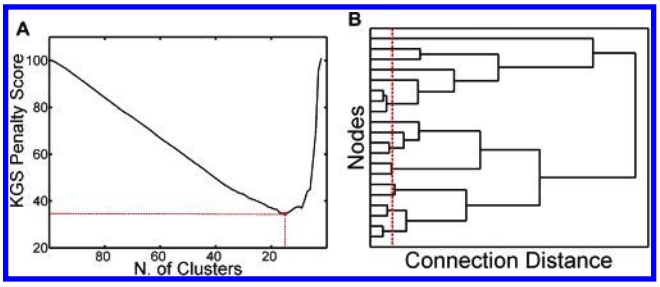

Figure 3. KGS function applied to the cluster outputs of the docking of SU4984 at the binding site of the FGFR1 tyrosine kinase domain. (A) Plot of the penalty values vs the clusters number. The dotted red line pinpoints the minimum penalty value. (B) The minimum KGS value corresponds to the dendrogram cutting point (dotted red line). Only the upper 20 nodes of the dendrogram are shown.

be performed to organize docking outputs. In the example, this was shown by studying H717 with both Dock (Figures 2E, F) and FlexX (Figures 2G, H). Finally, we can assess that the case of H* being less than 0.5 is not expected to occur in docking studies.

Cluster Analysis. Hierarchical−agglomerative cluster analyses were applied to the 57 (15 AutoDock outputs, 11 GOLD outputs, 15 Dock outputs, and 16 FlexX outputs) sets of docked poses, whose clusterability was earlier assessed by means of the H* test (Table 1). Remarkably, only for 48 out of 57 sets submitted to cluster analyses did docking software provide a pose with an RMSD less than 2.5 Å relative to the experimental conformation.

The hierarchical−agglomerative approach led to a nested sequence of partitions, and the KGS penalty score function was employed to select the functional one. For all of the 57 docking outputs, the curves obtained plotting the cluster numbers against the KGS penalty value displayed a very similar profile. For instance, in Figure 3A, the plot of the KGS cutting function applied to the clusters (using the average linkage rule, see below) of the outputs of the docking of SU4984 at the binding site of the FGFR1 tyrosine kinase domain (PDB code 1AGW) is reported.

In the left part of the plot, the high cluster number mainly contributed to the KGS penalty values and the function showed a slight negative slope. In the right part, the internal clustering spread was the driving force and the slope became steep and positive. The minimum of the KGS function showed the point at which the dendrogram had to be cut (see Figure 3B). Notably, around the KGS minimum, the function turned out to be very sensitive and the cutting point could always be univocally determined. The latest feature is a fundamental criterion for the choice of a cutting function. Moreover, thanks to the application of KGS, the cluster population turned out to be quite homogeneous.

Several approaches to tackle the issue of the cutting level selection are known.[25] Besides simple threshold rules, it is worth mentioning the work of Mojena,[26] who developed three cutting rules, the upper tail rule, the moving average quality control rule, and the double exponential smoothing rule, which rely upon intra- and intercluster distance distribution to select the cutting point. The major drawback of both simple threshold rules and more elaborate procedures based on intra- and intercluster distances is the reliance on parameters chosen by the investigator, carrying with them the possibility of deep influence from a priori expectations.

Table 2. Results for the Single Linkage Cluster Analysis Performed on the Outputs of the Four Docking Programs[a]

PDB code	NC	NNSC	RMSD Best	RMSD Pop	RMS NS	RMSD E	NC	NNSC	RMSD Best	RMSD Pop	RMSD NS	RMSD E
			Autodock						Gold			
1URW	23	5	0.87	1.74	1.25	1.49	21	5	1.91	2.44	2.44	3.58
1DI8	16	8	0.40	0.49	0.49	>5	7	4	0.61	0.95	0.95	1.33
1G5S			H*= 0.55						H* = 0.56			
1E1X	11	5	0.89	1.73	1.73	>5			H* = 0.58			
1AGW	15	9	1.44	>5	1.83	>5	9	6	0.72	>5	0.81	>5
1Q5K	16	9	0.89	1.26	1.26	3.70	11	7	0.90	2.51	1.42	1.88
1IEP	7	5	0.56	1.12	1.12	3.93	8	6	0.50	1.35	0.50	0.59
3ERT	15	5	1.15	1.68	1.68	2.03			H* = 0.54			
1S9P	17	11	0.19	1.12	0.2	1.17	18	10	0.2	1.16	0.24	1.64
1A28	18	9	0.40	0.54	0.44	0.47			H* = 0.51			
1EVE	10	7	0.93	2.05	2.05	>5	25	14	1.77	2.8	2.42	>5
1N5R	10	7	1.17	1.6	1.6	>5	23	14	>5	>5	>5	>5
1UT6	26	8	0.7	4.8	0.7	>5	17	7	0.90	0.97	0.97	4.06
1UV6	21	10	0.91	4.75	0.91	>5	27	13	1.77	3.82	1.89	4.04
1W51	12	5	1.91	3.5	3.5	2.4	15	8	0.71	1.03	1.03	2.71
1NNB	19	3	0.5	0.77	0.77	1.09			H* = 0.56			
			Dock						FlexX			
1URW	10	4	1.35	1.61	1.61	1.81	27	12	1.41	1.86	1.66	2.26
1DI8	13	3	>5	>5	>5	>5	17	15	2.38	>5	2.38	>5
1G5S	12	4	0.74	1.14	0.83	1.08	10	7	1.37	2.98	2.98	>5
1E1X	23	9	0.89	3.65	1.35	3.55	14	7	0.71	2.04	2.04	1.35
1AGW	15	8	1.20	1.23	1.23	>5	15	5	0.96	>5	1.16	0.96
1Q5K	26	10	0.69	0.69	0.69	1.65	16	11	1.4	>5	1.42	2.9
1IEP	28	9	>5	>5	>5	>5	7	5	1.00	>5	1.77	1.41
3ERT	21	10	0.85	1.07	1.07	2.02	23	10	1.26	1.71	1.51	2.2
1S9P	28	9	>5	>5	>5	>5	9	9	2.47	>5	2.63	2.53
1A28			H* = 0.58				10	10	>5	>5	>5	>5
1EVE	37	10	0.46	>5	0.6	3.73	10	9	0.93	2.05	2.05	3.12
1N5R	21	12	4.38	>5	>5	>5	12	9	>5	>5	>5	>5
1UT6	12	10	1.23	5.0	1.32	>5	18	12	>5	>5	>5	>5
1UV6	17	6	1.39	1.67	1.51	1.65	8	6	1.92	>5	1.93	2.00
1W51	17	6	2.15	2.57	2.57	2.40	14	9	3.4	>5	3.4	>5
1NNB	13	7	0.89	1.22	1.22	1.08	17	13	0.77	1.29	1.29	4.13

[a] RMSD values are calculated with respect to the experimental pose. NC = number of clusters; NNSC = number of nonsingleton clusters; RMSD Best = RMSD of the best docked pose; RMSD Pop = RMSD of the most populated cluster(s); RMSD NS = RMSD of the nonsingleton cluster closest to the experimental pose; RMSD E = RMSD of the best energy scoring pose.

In this context, KGS provides a simple and objective cutting method without a priori knowledge of the best possible solution.

In this study, the KGS cutting function was applied to the outputs of three different hierarchical−agglomerative clustering algorithms. In detail, single linkage,[23] average linkage,[23] and the Ward method[24] were applied to the 57 data sets for a total of 171 different clustering patterns. Every pattern was examined in terms of the number of clusters and the number of nonsingleton clusters (i.e., clusters with more than one element). The good intracluster homogeneity allowed us to consider the pose closest to the centroid as the one representative of the binding mode associated to each cluster. The representative poses were compared to the experimental one in terms of RMSD. The RMSD between the most populated cluster(s) and the experimental pose along with the RMSD of the nonsingleton cluster closest to the experimental pose were taken into account as comparative parameters. Moreover, the RMSD between the experimental pose and the best-docked one along with the RMSD between the best energy scoring pose and the experimental one were also computed.

In Table 2, the results of the cluster analysis with the single linkage rule are summarized. The number of clusters was quite low. Interestingly, several clusters comprised only one pose (singleton clusters or singletons), despite the fact that the KGS cutting function strongly promotes aggregation.[17]

In Table 3, the results of the cluster analysis with the average linkage rule are reported. Again, the number of clusters was quite low even though, as expected, it was higher than that obtained using the single linkage rule. Moreover, with the average linkage rule as well, several singletons were obtained.

Finally, in Table 4, the output results obtained applying the Ward method are shown. The number of clusters greatly increased relative to the average linkage results, while the number of singletons was quite reduced with respect to the other rules.

To better show the comparative results of the application of the clustering algorithms, in Figure 4, NU6027 (Figure 4A) docked with AutoDock at the CDK2 binding site (PDB code 1E1X) is reported.

In Figure 4B, a cluster obtained using the single linkage rule, which accounted for a population of 86 elements (out of 100), and comprising the pose closest to the experimental one is shown. The intracluster RMSD of the poses was 2.72 Å. Actually, the combination of the single linkage rule with the KGS penalty function strongly aggregated docking poses. This provided a very low number of clusters, which could however be quite inhomogeneous, as a consequence of the chaining effects of the single linkage rule.[23] In this respect, the application of the KGS cutting function could enhance

Table 3. Results for the Average Linkage Cluster Analysis Performed on the Outputs of the Four Docking Programs[a]

PDB code	NC	NNSC	RMSD Best	RMSD Pop	RMSD NS	RMSD E	NC	NNSC	RMSD Best	RMSD Pop	RMSD NS	RMSD E
			AutoDock						Gold			
1URW	23	6	0.87	1.74	1.74	1.49	16	11	1.91	2.44	2.44	3.58
1DI8	16	8	0.40	0.49	0.49	>5	18	12	0.61	0.85	0.85	1.33
1G5S			H*= 0.55						H* = 0.56			
1E1X	13	6	0.89	1.73	1.22	>5			H* = 0.58			
1AGW	15	10	1.44	>5	1.83	>5	9	6	0.72	>5	0.81	>5
1Q5K	20	12	0.89	0.89	1.05	3.70	16	11	0.90	2.51	0.90	1.88
1IEP	19	12	0.56	0.56	0.82	3.93	11	8	0.50	1.27	0.68	0.59
3ERT	12	7	1.15	1.52	1.52	2.03			H* = 0.54			
1S9P	16	11	0.19	1.12	0.2	1.17	19	10	0.2	1.16	0.24	1.64
1A28	11	11	0.40	0.55	0.4	0.47			H* = 0.51			
1EVE	20	11	0.93	1.82	1.27	>5	20	14	1.77	2.96	2.36	>5
1N5R	12	9	1.17	1.6	1.6	>5	18	9	>5	>5	>5	>5
1UT6	22	8	0.7	4.37	0.7	>5	20	12	0.90	0.98	0.98	4.06
1UV6	23	12	0.91	4.64	0.91	>5	24	13	1.77	3.82	1.88	4.04
1W51	18	13	1.91	3.36	3.36	2.4	14	9	0.71	1.03	1.03	2.71
1NNB	21	4	0.5	0.77	0.77	1.09			H* = 0.56			
			Dock						FlexX			
1URW	8	4	1.35	1.61	1.61	1.81	24	21	1.41	1.86	1.66	2.26
1DI8	11	3	>5	>5	>5	>5	17	17	2.38	>5	2.38	>5
1G5S	13	4	0.74	1.14	0.83	1.08	10	7	1.37	2.98	2.98	>5
1E1X	24	10	0.89	3.65	1.37	3.55	13	8	0.71	1.35	1.35	1.35
1AGW	16	9	1.20	1.23	1.23	>5	21	18	0.96	>5	0.96	0.96
1Q5K	17	16	0.69	1.04	0.8	1.65	16	13	1.4	>5	1.42	2.9
1IEP	27	11	>5	>5	>5	>5	7	7	1.00	>5	1.77	1.41
3ERT	19	15	0.85	1.09	1.09	2.02	25	20	1.26	1.51	1.51	2.2
1S9P	30	12	>5	>5	>5	>5	15	15	2.47	>5	2.63	2.53
1A28			H* = 0.58				14	12	>5	>5	>5	>5
1EVE	35	11	0.46	>5	0.6	3.73	20	11	0.93	1.82	1.26	3.12
1N5R	11	9	4.38	>5	4.66	>5	12	10	>5	>5	>5	>5
1UT6	9	9	1.23	5.00	1.27	>5	20	14	>5	>5	>5	>5
1UV6	15	6	1.39	1.67	1.55	1.65	14	12	1.92	>5	1.93	2.00
1W51	22	7	2.15	2.15	2.57	2.40	10	10	3.4	>5	3.4	>5
1NNB	16	6	0.89	1.35	0.89	1.08	15	13	0.77	1.45	0.87	4.13

[a] RMSD values are calculated with respect to the experimental pose. NC = number of clusters; NNSC = number of nonsingleton clusters; RMSD Best = RMSD of the best docked pose; RMSD Pop = RMSD of the most populated cluster(s); RMSD NS = RMSD of the nonsingleton cluster closest to the experimental pose; RMSD E = RMSD of the best energy scoring pose.

the chaining effects of the single linkage rule. When average linkage was used, the same 86 poses corresponded to three clusters (Figure 4C). Indeed, fewer chaining effects were identified, as clearly shown by comparing the results of Figure 4B with those of Figure 4C. With the average linkage rule, the cluster comprising the pose closest to the crystallographic one was composed of six docking conformations, whereas the other two by were composed of 79 and 1, respectively. The RMSD among the six docking conformations was as low as 1.4 Å. Interestingly, the pose closest to the experimental one was not included in the most populated cluster, being however within a nonsingleton. When the Ward method (Figure 4D) was used, the 86 poses corresponded to five clusters, and that comprising the docking conformation closest to the crystallographic binding mode was composed of only three poses. The internal RMSD among the three poses was 0.8 Å. By this comparison, it is clear that, while the single linkage rule strongly aggregated docking poses and in turn showed chaining effects, the Ward method provided very space-conservative clusters. However, the high cluster number obtained with the Ward algorithm could provide unmanageable results. Indeed, the KGS function applied to the average linkage outputs was a good compromise between a reasonably low number of clusters and internal homogeneity.

From an inspection of Tables 2−4, it can be argued that a straight correlation was identified neither between the most populated clusters and the best-predicted poses nor between the most populated clusters and the best energy score poses. Even if it would have been appealing to provide a correlation between the most populated clusters and putative binding modes (this occurred in 29 out of 48 cases, in which docking programs were able to provide the experimental pose and a cluster analysis was actually performed), such a relationship was not achieved. However, the experimental pose was always within nonsingleton clusters, and accordingly, singletons were always real outliers, as a consequence of the great aggregating behavior of a clustering approach based on the combination of the KGS penalty function and linkage algorithms. Furthermore, we noticed that the most populated cluster often was just slightly bigger than other significant clusters (Table 5), since more than a single binding mode might be likely for some ligand−protein complexes.[27,28] In Table 5, the significant clusters obtained using the three different algorithms are reported.

In statistical terms, significant clusters are the ones whose populations depart by more than 2σ from the mean cluster population. As expected, the Ward method very often did not provide even a single significant cluster and, in other cases, provided more than one, whereas the single and average linkage rules seldom provided more than a single significant cluster.

In light of the above considerations, we conclude that the combination of the average linkage rule and the KGS cutting function may provide a new and promising clustering approach to organize the high number of docking poses and to improve the reliability of molecular docking results. This allows an objective and strictly mathematical partition of

Table 4. Results for the Ward Method Cluster Analysis Performed on the Outputs of the Four Docking Programs[a]

PDB code	NC	NNSC	RMSD Best	RMSD Pop	RMSD NS	RMSD E	NC	NNSC	RMSD Best	RMSD Pop	RMSD NS	RMSD E
			AutoDock						Gold			
1URW	27	12	0.87	1.13	1.13	1.49	20	12	1.91	2.39	2.33	3.58
1DI8	21	13	0.40	0.67	0.53	>5	22	22	0.61	0.76	0.76	1.33
1G5S			H*= 0.55						H*= 0.56			
1E1X	14	9	0.89	1.8	1.6	>5			H* = 0.58			
1AGW	15	10	1.44	>5	1.83	>5	9	6	0.72	>5	0.81	>5
1Q5K	19	17	0.89	1.27	1.25	3.70	17	16	0.90	1.86	0.90	1.88
1IEP	12	11	0.56	0.82	0.82	3.93	13	11	0.50	1.27	0.58	0.59
3ERT	13	9	1.15	1.7	1.4	2.03			H* = 0.54			
1S9P	15	12	0.19	1.12	0.2	1.17	21	12	0.2	1.18	0.24	1.64
1A28	12	12	0.40	0.54	0.4	0.47			H* = 0.51			
1EVE	23	16	0.93	1.25	1.25	>5	20	17	1.77	2.96	2.96	>5
1N5R	26	18	1.17	1.80	1.69	>5	10	9	>5	>5	>5	>5
1UT6	22	10	0.7	1.88	0.7	>5	20	15	0.90	1.22	0.98	4.06
1UV6	10	9	0.91	1.33	1.33	>5	26	18	1.77	3.88	1.89	4.04
1W51	26	21	1.91	2.6	2.6	2.4	23	14	0.71	0.71	0.71	2.71
1NNB	22	11	0.5	0.96	0.6	1.09			H* = 0.56			
			Dock						FlexX			
1URW	16	9	1.35	1.52	1.39	1.81	21	20	1.41	2.08	1.51	2.26
1DI8	11	4	>5	>5	>5	>5	15	15	2.38	>5	2.60	>5
1G5S	12	8	0.74	1.1	0.83	1.08	16	13	1.37	1.6	1.6	>5
1E1X	15	11	0.89	3.65	1.42	3.55	16	14	0.71	0.81	0.81	1.35
1AGW	16	11	1.20	1.28	1.26	>5	3	3	0.96	10.38	1.16	0.96
1Q5K	19	19	0.69	0.8	0.8	1.65	15	15	1.4	>5	1.42	2.9
1IEP	22	15	>5	>5	>5	>5	12	12	1.00	1.77	1.77	1.41
3ERT	19	19	0.85	1.09	1.06	2.02	16	16	1.26	1.51	1.51	2.2
1S9P	34	19	>5	>5	>5	>5	14	14	2.47	>5	2.63	2.53
1A28			H* = 0.58				16	16	>5	>5	>5	>5
1EVE	28	16	0.46	>5	0.6	3.73	23	17	0.93	1.25	1.25	3.12
1N5R	16	14	4.38	>5	4.66	>5	15	11	>5	>5	>5	>5
1UT6	9	9	1.23	>5	1.5	>5	22	15	>5	>5	>5	>5
1UV6	9	5	1.39	1.62	1.51	1.65	13	13	1.92	>5	1.93	2.00
1W51	23	11	2.15	2.51	2.3	2.40	11	11	3.4	>5	3.4	>5
1NNB	16	14	0.89	0.97	0.93	1.08	21	20	0.77	1.45	0.87	4.13

[a] RMSD values are calculated with respect to the experimental pose. NC = number of clusters; NNSC = number of nonsingleton clusters; RMSD Best = RMSD of the best docked pose; RMSD Pop = RMSD of the most populated cluster(s); RMSD NS = RMSD of the nonsingleton cluster closest to the experimental pose; RMSD E = RMSD of the best energy scoring pose.

outputs of reiterated docking runs, which are required for any accurate docking simulation.

Cluster analyses were already implemented in docking programs. For instance, the genetic-algorithm-based suites that by definition provide a population of output poses implement postprocessing clustering approaches. For instance, the AutoDock cluster analysis is carried out by looping through the conformations, comparing the RMSDs, and grouping together elements, whose RMSDs are within an arbitrarily chosen threshold value. Some remarks seem somewhat relevant. First, the final result is not invariant with respect to the initial order of the elements, and thus, the first element is always a cluster leader. Consequently, the first clusters generally become wider and less space-conservative than the last ones. Furthermore and very importantly, different threshold values lead to different partitioning patterns. As a consequence, the AutoDock cluster analysis requires high-level user intervention and clearly does not match the basic requirements of an accurate clustering approach. GOLD and FlexX implement a complete linkage[23] clustering algorithm as a postprocessing tool. The complete linkage approach tends to find compact clusters of equal diameter and is by definition not space-conservative. Moreover, while GOLD does not provide any criterion to select a functional partition, FlexX implements the RMSD-based threshold distance as a cutting rule. Again, high-level user intervention might greatly affect the accuracy of the partitioning results.

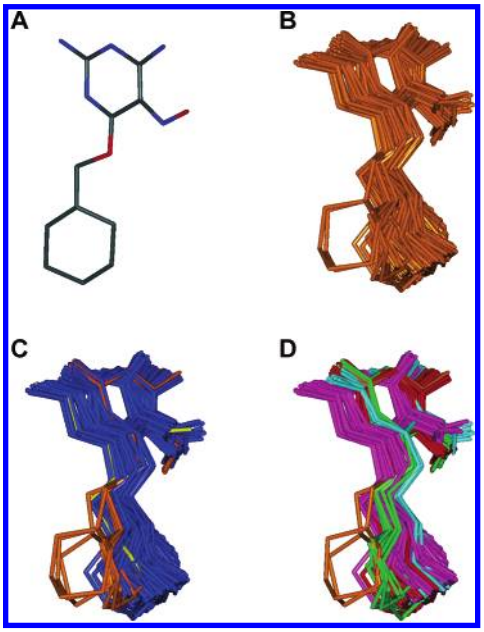

Figure 4. NU6027 (A) docked with the AutoDock program at the CDK2 binding site. The cluster comprising the pose closest to the experimental one is shown in orange. (B) Single linkage cluster rule results. The main cluster accounts for 86 poses. (C) The same 86 poses reported in B, correspond to three different clusters (79 poses in blue, 1 in pose yellow, and 6 in poses orange) using the average linkage rule. (D) The same 86 poses reported in B correspond to five different clusters (44 poses in magenta, 18 in poses red, 14 in poses cyan, 7 in poses green, and 3 in poses orange) using the Ward method.

Table 5. Significant Clusters Obtained Using the Three Different Algorithms[a]

	single	average	Ward	single	average	Ward
	AutoDock			GOLD		
1URW	1(36) 3(22)	5(39)	5(16)	10(48)	1(34)	9(15)
1DI8	5(63)	5(62)	5(23)	7(72)	13(22)	17(9)
1G5S	H* = 0.55	H* = 0.55	H* = 0.55	H* = 0.56	H* = 0.56	H* = 0.56
1E1X	6(86)	3(73)	12(44)	H* = 0.58	H* = 0.58	H* = 0.58
1AGW	13(33)	–	–	5(46)	7(46)	–
1Q5K	7(46)	15(28)	4(13) 7(12)	5(42)	14(42)	–
1IEP	3(52)	14(43)	10(40)	4(65)	–	–
3ERT	8(57)	2(32)	–	H* = 0.54	H* = 0.54	H* = 0.54
1S9P	6(40)	9(40)	14(40)	3(30)	3(30)	10(16)
1A28	17(49)	5(26)	10(22)	H* = 0.58	H* = 0.58	H* = 0.58
1EVE	9(77)	3(56)	18(18)	14(25)	10(20)	12(14)
1N5R	1(36)	3(36)	16(15) 17(19)	5(39)	16(21)	–
1UT6	1(42)	3(24) 7(27)	10(24)	13(51)	1(28)	15(14)
1UV6	12(28)	17(19)	7(27)	15(25)	12(25)	14(17) 26(15)
1W51	3(83)	5(36)	1(10)	5(34)	10(34)	11(14)
1NNB	5(80)	1(77)	6(30)	H* = 0.56	H* = 0.56	H* = 0.56
	Dock			FlexX		
1URW	3(79)	3(80)	11(30)	1(30)	7(14) 19(12)	–
1DI8	1(86)	1(68)	1(42)	11(20)	11(20)	7(20)
1G5S	6(72)	3(71)	–	3(50)	3(50)	–
1E1X	15(36)	13(36)	11(36)	1(66)	8(56)	6(16) 12(16)
1AGW	3(35) 5(33)	8(35)	6(25) 7(27)	9(80)	7(23)	3(92)
1Q5K	6(40)	5(19)	–	1(34)	7(29)	5(20)
1IEP	3(33) 26(31)	7(20) 12(33)	15(21) 17(20)	1(66)	–	–
3ERT	1(21) 12(20)	17(16)	19(12)	5(34) 12(30)	7(13)	15(14)
1S9P	6(55)	1(31) 5(20)	9(11) 10(13) 13(10)	3(57)	4(21)	14(23)
1A28	H* = 0.58	H* = 0.58	H* = 0.58	5(26)	9(24)	–
1EVE	2(23) 7(19)	4(23) 13(19)	8(23) 25(19)	9(77)	3(56)	18(18)
1N5R	12(39) 17(25)	1(31)	10(17)	–	–	–
1UT6	12(39)	9(39)	–	17(41)	19(41)	1(16) 16(19)
1UV6	11(60)	5(55)	–	5(42)	8(21)	5(17)
1W51	3(73)	7(70)	5(28)	10(32)	3(32)	–
1NNB	8(58)	5(35)	–	5(58)	7(28)	–

[a] The population of the significant clusters is reported in parenthesis. " – " means that no significant clusters were found.

Consensus Clustering. Docking performances of different tools are known to change dramatically depending on the chemical nature of the ligand and the physicochemical traits of the active site.[7] None of the four docking tools employed in this study significantly outperformed the others, even though genetic algorithms were able to achieve a slightly better accuracy, as previously reported by Rognan and coworkers.[10,29]

In light of these results, we applied a consensus approach carrying out on each complex the best-performing clustering procedure (namely, average linkage with the KGS function; see above) for the complete output of AutoDock, GOLD, FlexX, and Dock (400 poses). In Table 6, the results of the consensus clustering are reported. As expected, in most of the cases, the total number of clusters was lower than the sum of the clusters obtained from the analysis separately carried out on the outputs of each docking run (see above). This showed that the conformational space explored by the four docking programs widely overlapped, namely, poses generated by different docking programs belonged to the same cluster.

Table 6. Results for the Consensus Clustering Performed on the Outputs of the Four Docking Programs[a]

PDB ID	GOLD	AUTO	FLEXX	DOCK	NC	SC	POP	RMSD
1URW	3.58	1.49	2.26	1.81	51	Cluster01	172	1.6
						Cluster12	62	1.97
1E1X	1.1	>5	1.35	3.55	41	Cluster04	77	2.44
						Cluster11	108	0.94
1G5S	0.73	1.48	>5	1.08	50	Cluster03	61	0.77
						Cluster04	138	1.10
1DI8	1.33	>5	>5	>5	38	Cluster05	135	0.47
						Cluster11	112	>5
1AGW	>5	>5	0.96	>5	50	Cluster11	40	1.19
						Cluster23	69	>5
						Cluster24	43	>5
1Q5K	1.88	3.7	2.9	1.65	52	Cluster30	213	1.35
1IEP	0.59	3.93	1.41	>5	85	Cluster05	149	0.86
3ERT	1.68	2.03	2.2	2.02	23	Cluster18	79	1.38
						Cluster05	68	1.70
1S9P	1.64	1.17	2.53	>5	57	Cluster32	157	1.12
1A28	0.27	0.47	>5	0.56	60	Cluster05	102	0.53
						Cluster06	198	0.33
1EVE	>5	>5	3.73	3.12	53	Cluster03	45	>5
						Cluster11	38	2.79
						Cluster15	58	1.82
1N5R	>5	>5	>5	>5	33	Cluster15	52	>5
						Cluster19	85	>5
1UT6	4.06	>5	>5	>5	51	Cluster26	98	1.15
						Cluster04	57	>5
1UV6	4.04	>5	2.00	1.65	55	Cluster43	107	1.62
1W51	2.71	2.4	>5	2.4	70	Cluster08	124	2.40
						Cluster40	70	>5
1NNB	0.75	1.09	4.13	1.08	53	Cluster09	182	0.75

[a] RMSD values are calculated with respect to the experimental pose. NC = number of clusters; SC = number of significative clusters; POP = population of significant cluster; RMSD = RMSD value of the representative pose from each cluster. For every docking software example, the RMSD value of the best energy pose according to the scoring function is reported.

Remarkably, each consensus analysis always provided at least one significantly (in statistical terms) populated cluster and, more importantly, never more than three. In 15 out of the 16 complexes here investigated, the docking conformation resembling the crystal structure (RMSD less than 2.5 Å) was the representative pose related to one of the statistically significant clusters. This means that, for almost all (15/16) of the cocrystals here studied, analyzing up to three docking poses would have always allowed identification of the correct binding mode.

The only crystallographic pose not identified by means of the here-described consensus clustering was that of the complex between acetylcholinesterase (AChE) and propidium. However, propidium, being a surface-binding ligand, has been shown by both experiments and computations to be able to bind at the AChE peripheral site in at least two different modes.[27,28]

Eventually, by an inspection of Table 6, it can be seen that none of the exploited docking software was able to provide more accurate results than those achieved by using consensus clustering. Conversely, applying our approach, one has often to investigate more than one single pose, this number being, however, on the order of a few units.

In summary, the consensus clustering procedure allowed us to (i) reduce the actual number of poses to be investigated, (ii) greatly enhance the accuracy of the proposed clustering approach, and (iii) find in most (15/16) of the investigated

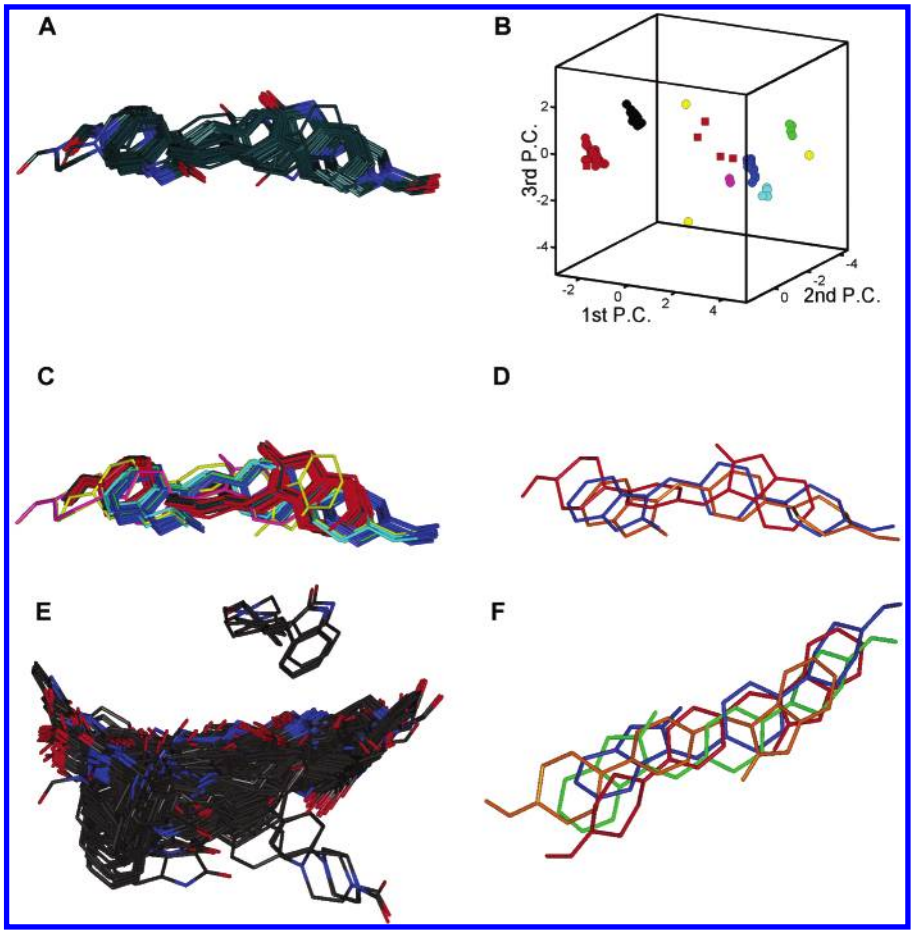

Figure 5. (A) Ensemble of poses resulting from the study of 1AGW with GOLD. (B) H* test results plotted against three principal components. (C) Application of the average linkage cluster analysis. The six nonsingleton clusters (46 poses in red, 21 poses in dark gray, 20 poses in blue, 4 poses in green, 4 poses in cyan, and 2 in poses magenta) are shown. (D) The crystallographic pose (orange), the best energy pose (red, RMSD from the crystal 10.66 Å), and the representative pose of the cluster closest to the crystallographic solution (blue, RMSD from the crystal 0.81 Å) are reported. (E) Ensemble of poses resulting from the study of 1AGW with AutoDock, GOLD, FlexX, and DOCK. (F) The crystallographic pose (orange) and the poses representative of the three significantly populated clusters (Cluster11, pose 350, RMSD from the crystal 1.19 Å, red; Cluster23, pose 41, RMSD from the crystal 9.94 Å, blue; Cluster24, pose 313, RMSD from the crystal 9.97 Å, green) are reported.

complexes the pose closest to the experimental one within a statistically significant cluster.

CONCLUSIONS AND OUTLOOK

In this paper, we have presented a theoretical study aimed at comparing different hierarchical−agglomerative clustering algorithms to organize docking outputs and to improve the reliability of simulation results. Actually, a well-known open issue in docking studies is that reiterated runs, which are required for an accurate protocol, provide several different docking poses for the same ligand−target complex. The docking users have to therefore face the problem of choosing a single binding mode from a burden of several different docking conformations. Often, scoring functions might not help with the choice.[30] To tackle this issue, we proposed the use of a clustering approach.

To summarize, in Figure 5, the results of the present integrated clustering approach applied to the study of 1AGW with GOLD are reported. First, the clusterability of the docking ensemble (Figure 5A) was assessed. Here, we showed that a modified version of the Hopkins test was very well suited to this aim. In Figure 5B, the data plotted against three principal components, accounting for more than 90% of the variance, are reported. The H* was close to 1 (equal to 0.91), as a consequence of the fact that virtual points (red squares in the figure) were randomly placed in empty regions around possible clusters. One may conclude that such a data distribution clearly exhibits a predisposition to clustering. In this case, a cluster analysis has to be carried out. We propose that the combination of the KGS penalty function with the average linkage rule is the best integrated approach to partition docking poses. KGS provides a functional classification in a strictly mathematical manner, while the average linkage algorithm has been here-demonstrated to be a good compromise between a low number of clusters and space-conservative results. Actually, while the single linkage algorithm showed chaining effects, the Ward method provided a high number of nonsignificant clusters. In Figure 5C, poses representative of the six nonsingleton clusters are reported using the same color code of Figure 5B. Interestingly, the most populated cluster (46 poses, red in Figure 5C) comprised the best energy pose but not that closest to the experimental one. In Figure 5D, the crystallographic pose (orange), the best energy pose (red, RMSD from the crystal equal to 10.66 Å), and the representative pose of the cluster (20 poses) closest to the experimental solution (blue, RMSD from the crystal equal to 0.81 Å) are reported. Singletons

again were the actual outliers. Finally, aimed at further improving the accuracy of our procedure, we carried out a consensus clustering. All docking outputs (400 poses, see Figure 5E) coming from the four programs were clustered using average linkage and the KGS penalty function. Actually, we increased the accuracy of our results, with as many as 15 out of 16 investigated cocrystals correctly being reproduced following this approach, greatly reducing the number of clusters to study. For instance, in Figure 5F, three statistically significant clusters of 1AGW are reported. As shown in the picture, the pose closest to the experimental one (orange in Figure 5F) was within a statistically significant cluster. In contrast, for the clustering applied to the GOLD outputs (i.e., the best-performing docking algorithm for this cocrystal), all the nonsingletons (six, Figure 5C) had to be considered to identify the pose closest to the experimental one (Figure 5D).

In conclusion, the present approach allowed us to univocally point to singleton clusters as outliers and to greatly reduce the number of significant poses (nonsingletons) among which the experimental ones were always present. Moreover, using a consensus clustering procedure, for 15 out of 16 cocrystals here investigated, the pose closest to the experimental one was within a statistically significant cluster, whose number was always between 1 and 3. This means that analyzing few docking poses likely allows one to identify the correct one.

Thanks to the easy automation of the present integrated clustering approach, we propose its possible implementation for partitioning outputs from reiterated runs of virtual screening of databases of ligands. In this respect, we believe that clustering methods would improve the accuracy of virtual screening results. Finally, conformational analysis of flexible organic molecules or peptides might be another field where such a clustering approach may help the organization of computational results. Actually, a threshold distance-based clustering approach was earlier shown to greatly help the rationalization of outputs of conformational search methods.[31] This approach, first implemented in the MacroModel 5.5[32] suite, is nowadays widely applied in partitioning outputs of sampling algorithms such as Monte Carlo and molecular dynamics simulations.

ACKNOWLEDGMENT

We thank W. Rocchia for very useful discussions. We also thank the anonymous reviewers for the very useful suggestions that greatly improved the quality of the present paper. MIUR-COFIN2004 and FIRB are gratefully acknowledged for financial support.

Supporting Information Available: The complete list of the crystallographic complexes employed in the present study is reported in Table 1SI. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Yokoyama, S. Protein expression systems for structural genomics and proteomics. *Curr. Opin. Chem. Biol.* **2003**, *7*, 39−43.
(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(4) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(5) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.
(6) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(7) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(8) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(9) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.
(10) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(11) Simonson, T.; Archontis, G.; Karplus, M. Free energy simulations come of age: protein−ligand recognition. *Acc. Chem. Res.* **2002**, *35*, 430−437.
(12) Gervasio, F. L.; Laio, A.; Parrinello, M. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.* **2005**, *127*, 2600−2607.
(13) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.
(14) Chema, D.; Eren, D.; Yayon, A.; Goldblum, A.; Zaliani, A. Identifying the binding mode of a molecular scaffold. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 23−40.
(15) Kallblad, P.; Mancera, R. L.; Todorov, N. P. Assessment of multiple binding modes in ligand-protein docking. *J. Med. Chem.* **2004**, *47*, 3334−3337.
(16) Hopkins, B. A new method for determining the type of distribution of plant individuals. *Ann. Bot. (Oxford, U. K.)* **1954**, *18*, 213−227.
(17) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. *Protein Eng.* **1997**, *10*, 737−741.
(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. P. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(19) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(20) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(21) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(22) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(23) Everitt, B. S.; Landau, S.; Leese, M. *Cluster analysis*; Arnold, a member of the Hodder Headline Group: London, 2001.
(24) Ward, J. H. J.; Hook, M. E. Application of a hierarchical grouping procedure to problem of grouping profiles. *Educ. Psychol. Meas.* **1963**, *23*, 69−92.
(25) 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.
(26) Mojena, R. Hierarchical grouping methods and stopping rules: an evaluation. *Comput. J.* **1977**, *20*, 353−363.
(27) Bourne, Y.; Taylor, P.; Radic, Z.; Marchot, P. Structural insights into ligand interactions at the acetylcholinesterase peripheral anionic site. *EMBO J.* **2003**, *22*, 1−12.
(28) Cavalli, A.; Bottegoni, G.; Raco, C.; De Vivo, M.; Recanatini, M. A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase. *J. Med. Chem.* **2004**, *47*, 3991−3999.
(29) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins* **2002**, *47*, 521−533.
(30) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein−ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.
(31) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(32) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R. M. J.; Lipton, M. A.; Caulfield, C. E.; Chang, G.; Hendrickson, T. F.; Still, W. C. MacroModel − an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *1*, 440−467.

CI050141Q