J Mol Model (2009) 15:551–560
DOI 10.1007/s00894-008-0360-6

SOFTWARE REPORT

3-D clustering: a tool for high throughput docking

John P. Priestle

Received: 15 April 2008 / Accepted: 18 September 2008 / Published online: 16 December 2008

Abstract This report describes a computer program for clustering docking poses based on their 3-dimensional (3D) coordinates as well as on their chemical structures. This is chiefly intended for reducing a set of hits coming from high throughput docking, since the capacity to prepare and biologically test such molecules is generally far more limited than the capacity to generate such hits. The advantage of clustering molecules based on 3D, rather than 2D, criteria is that small variations on a scaffold may bring about different binding modes for molecules that would not be predicted by 2D similarity alone. The program does a pose-by-pose/atom-by-atom comparison of a set of docking hits (poses), scoring both spatial and chemical similarity. Using these pair-wise similarities, the whole set is clustered based on a user-supplied similarity threshold. An output coordinate file is created that mirrors the input coordinate file, but contains two new properties: a cluster number and similarity to the cluster center. Poses in this output file can easily be sorted by cluster and displayed together for visual inspection with any standard molecular viewing program, and decisions made about which molecule should be selected for biological testing as the best representative of this group of similar molecules with similar binding modes.

Keywords Binding mode · Chemical similarity · Clustering · Docking

J. P. Priestle (✉)
Center for Proteomic Chemistry,
Novartis Institutes for Biomedical Research,
4002 Basel, Switzerland
e-mail: john.priestle@novartis.com

Introduction

Because the number of hits coming out of a high throughput docking (HTD) analysis usually far exceeds the number of compounds that can actually be biologically tested, various filtering procedures are used to reduce the hit list to attain a much smaller number of compounds that best represents the whole set. One of the most valuable procedures is the clustering of molecules, usually based on 2-dimensional (2D) chemical descriptors, which allows the selection of a few representative compounds from groups of related molecules. There are a number of algorithms and programs for performing this [1–7]. 2D clustering is based on the assumption that similar molecules bind to the target protein in a similar manner. While generally a reasonable assumption, there are cases of closely related molecules binding to proteins in surprisingly different ways (e.g., [8–10]). One attempt to address this issue for docked poses was the generation of structural interaction fingerprint (SIFt) patterns [11], which describe the interactions between a ligand and the protein to which it is bound. These can then be clustered by standard fingerprint clustering methods. These were first used to cluster multiple potential binding modes of a single ligand generated by docking, although the method is easily extended to be used with docking poses from different ligands against the same protein target.

This report describes a computer program that compares docked molecules based on the docking poses themselves, rather than the target protein to which they bind and is like 2D clustering, except that it is based both on chemical and spatial similarity. Throughout this report a distinction will be made between “compound” (or “molecule”), which refers to a 2D structure, atom connection table or SMILES string and “pose” which refers to a 3D molecular structure with a defined conformation and coordinates in 3D space, for example, after being docked to a protein.

Methods

CLSTR3D is a computer program written in FORTRAN. As input it requires a set of pose coordinates in .sdf format and, optionally, the protein structure, in PDB format, to which the compounds were docked. The molecules being clustered are typically poses from a high throughput docking (HTD) experiment and usually represent the single best fit of each molecule to a specific target protein and therefore are all in a common 3D reference frame. No fitting (optimization of molecular overlap) is done to the poses prior to comparison. The output file mirrors the input file except that two new properties per pose are added: “cluster_number” and “cluster_similarity”, the number of the cluster that the pose belongs to and its similarity to the cluster center, respectively.

Chemical similarity

An atomic similarity matrix was created based on the chemically advanced template search (CATS) fingerprint developed for scaffold hopping [12]. The functional class considers whether an atom is lipophilic, an H-bond acceptor, an H-bond donor, negatively charged, and/or positively charged. Both the charge and existence of attached hydrogen atoms of hydrophilic atoms determine their CATS functional class, so all possibilities were defined. Examination of vendor catalogs of “drug-like” or “lead-like” molecules (1.7 million unique compounds) indicated that such molecules are almost exclusively composed of relatively few elements: in order of abundance H, C, N, O, S, F, Cl, Br, P, I, and Si with other elements accounting for less than 0.0002% of the atoms examined, mostly boron and metals. In the interest of speed, it was decided to ignore hydrogen atoms. Although the hydrogen atoms themselves are not examined, their presence is still carried by the classification of the non-hydrogen atom to which they were attached. Seven functional class could be defined (Table 1).

No cases of hydrophilic atoms with a negative charge and attached hydrogen atoms were found among the 80 million atoms examined. Hydrophobic atoms with a charge (very rare, but extant) are no longer considered to be lipophilic. For the purpose of functional class determination, both oxygen atoms of a deprotonated carboxylic acid are considered negatively charged. This is also true for other negatively charged groups where the charge is distributed across two or more oxygen atoms (e.g., nitrate, sulphonate, phosphinate, etc.).

A functional class similarity matrix was created, where the similarity between functional classes was scored according to the number of shared fingerprint properties (Table 2). To highlight exact functional class matches and down weight more distant relationships, a score of 2^{sfb}, where “sfb” is the number of shared fingerprint bits between the two functional class, was used.

Comparing 3D structures

Since the structures being compared typically come from docking the compounds to the same protein target, the structures are already in a common 3D reference frame. All structures are compared pair-wise, atom by atom. An atom from structure *j* closest to a specific atom from structure *i* is considered matched if their separation is less than 1Å. In the rare case where more than one atom in structure *j* is closer than 1Å from the atom in structure *i*, only the closest atom is considered. The score for the matched atoms will be

$$\mathrm{wt_{ij}}\left(\mathrm{SM}\left(\mathrm{FC_i},\mathrm{FC_j}\right)\times\left(1.0-\mathrm{dist}^3\right)\right) \tag{1}$$

Where $\mathrm{SM(FC_i,FC_j)}$ is the functional class similarity matrix value (Table 2) between the functional class of atom

Table 1 Functional class definitions

FC	Lipo	+	−	HB Acc	HB Don	description
1	1	0	0	0	0	lipophilic (C/Si/S/halogens)
2	0	0	0	1	0	hydrophilic, uncharged, without hydrogen atoms
3	0	0	0	1	1	hydrophilic, uncharged, with hydrogen atoms
4	0	0	1	1	0	hydrophilic, negative, without hydrogen atoms
5	0	1	0	0	0	hydrophilic, positive, without hydrogen atoms
6	0	1	0	0	1	hydrophilic, positive, with hydrogen atoms
7	0	0	0	0	0	phosphorus, hexavalent sulfur

Functional classes (FC) are based on the following atomic properties: lipophilicity (lipo), charge (+/−), hydrogen bond acceptor (HB_Acc), and hydrogen bond donor (HB_Don), where “1” indicates that the functional class possesses this property and “0” indicates that it does not. Note that only the elements H, C, N, O, S, P, F, Cl, Br, I, and Si are acceptable, which covered >99.9998% of the 1.7 million drug-like and lead-like vendor compounds (80 million atoms) examined, although only 94.8% of the NCI collection. All such possible functional classes are covered except negative, hydrophilic with hydrogen atoms, for which no examples were found.

Table 2 Atomic similarity matrix of functional classes

FC	1	2	3	4	5	6	7	FC description
1	32	8	4	4	8	4	16	lipophilic (C/Si/S/halogens)
2	8	32	16	16	8	4	16	hydrophilic, uncharged, without hydrogen atoms
3	4	16	32	8	4	8	8	hydrophilic, uncharged, with hydrogen atoms
4	4	16	8	32	4	2	8	hydrophilic, negative, with hydrogen atoms
5	8	8	4	4	32	16	16	hydrophilic, positive, without hydrogen atoms
6	4	4	8	2	16	32	8	hydrophilic, positive, with hydrogen atoms
7	16	16	8	8	16	8	32	phosphorus, hexavalent sulfur

Each similarity element = 2^{sfb} , where *sfb* = number of *s*hared *f*ingerprint *b*its between the two functional classes).

i (FC_i) and functional class of atom *j* (FC_j), "dist^3" is the cube of the distance between the atoms and wt_{ij} is an optional weighting factor based on the interaction of atoms *i* and *j* with the protein (see below). 1.0 is the cube of the maximum distance (1Å). The cube of the distance is used since the probability of finding an atom purely by chance increases by the volume of the sphere defined by the search distance. If no atom in structure *j* is found within 1Å of the examined atom in structure *i*, a score of 0 is assigned. The atom-by-atom scores are summed and divided by the maximum possible score (= $\Sigma wt_{ij} \times 32 \times$ number of atoms in structure *i*) to get a 3D similarity score. Such a score can range from 1.0 (perfect match in 3D space and functional class) to 0.0 for no atom in structure *j* found within 1Å of any atom in structure *i*. In this way, an N × N matrix of similarity comparisons is built up, where N is the number of poses being compared. It should be noted that only identical structures can give similarities of 1.0. Poses that match perfectly in functional class will not match perfectly spatially because of differences in bond lengths. It should also be noted that comparing pose A to pose B will not necessarily give the same score as comparing pose B to pose A. In particular, if molecule A is a substructure of molecule B and overlaps perfectly with that part of molecule B in 3D, then comparing A to B will give a 3D similarity of 1.0, whereas comparing B to A will give a far lower score, since some of the atoms of pose B will have no corresponding atoms in pose A. Therefore, when comparing two poses, the larger of the two will be compared to the smaller and the results placed in cells (i,j) and (j,i,). If the two poses are the same size (identical number of non-hydrogen atoms), then only the first pose will be fit to the second one. This cuts the matrix calculation time by roughly half, generates a symmetrical square matrix and penalizes comparisons of molecules of vastly different sizes, with the advantage that the final clusters are more likely to contain molecules of similar size. A simple example of a 3D similarity calculation is worked out in Fig. 1 and Table 3.

Weights can be applied to the individual atoms based on the number of interactions they make with the protein. A simple counting of interactions is done; no attempt is made to estimate the strength of the interaction. If the distance between a protein atom and a ligand atom is less than 1.2× their combined van der Waals radii, an interaction is counted. The 20% increase in van der Waals radii is to compensate for possible coordinate error and for hydrogen atoms that may be attached to the atoms examined, since only non-hydrogen atoms are examined. Weighting may be desired since atoms interacting with the protein are more important to binding than those exposed to solvent. The weight applied is [number of protein interactions + 1], so that all atoms are still considered, but weighted differently. When comparing two structures, the total weight given to an atomic comparison is the sum of the weights of both matching atoms. If no matching atom is found, the weight is twice the weight of the search atom. Only non-hydrogen protein atoms are considered. Solvent and other non-protein molecules are ignored.

Clustering

Since all similarity comparisons have been pre-calculated, singletons can be quickly identified as poses with no similarity score above the user-defined threshold. They are flagged as such and removed from further consideration. The clustering algorithm employed is similar to the quality threshold cluster method developed by Heyer et al. for gene clustering [13]. Cluster centers are found by examining the pose similarity matrix for the pose with the largest number of "similar" structures among the other structures, i.e., structures with similarity scores greater than the user-

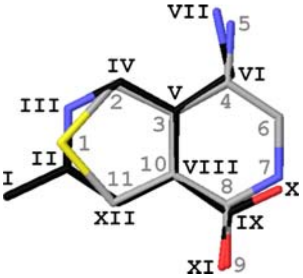


Fig. 1 Spatial comparison of two docked molecules: molecule A (gray carbon atoms, Arabic numerals) and molecule B (black carbon atoms, Roman numerals). Scoring their 3D similarity atom-by-atom is illustrated in Table 3

Table 3 Similarity calculation example

Molecule A			Molecule B						partial	max	atom
Atom	Element	FC	Atom	Element	FC	Dist.	$1.0\text{-}d^3$	$SM_{(B,A)}$	score	score	Score
1	S	1	II	C	1	0.641	0.737	32	23.6	32	0.737
2	C	1	IV	C	1	0.166	0.995	32	31.8	32	0.995
3	C	1	V	C	1	0.043	1.000	32	32.0	32	1.000
4	C	1	VI	C	1	0.199	0.992	32	31.7	32	0.992
5	N	3	VII	N	6	0.429	0.921	8	7.4	32	0.230
6	C	1	–	–	–	>1.0	0.000	0	0.0	32	0.000
7	N	3	X	O^-	4	0.298	0.974	8	7.8	32	0.244
8	C	1	IX	C	1	0.180	0.994	32	31.8	32	0.994
9	O	2	XI	O^-	4	0.123	0.998	16	16.0	32	0.499
10	C	1	VIII	C	1	0.028	1.000	32	32.0	32	1.000
11	C	1	XII	C	1	0.124	0.998	32	31.9	32	0.998
Total									246.0	352	0.699

Similarity of molecule B (black carbon atoms, Roman numerals) to molecule A (gray carbon atoms, Arabic numerals) from Fig. 1 without protein interaction weighting. The "atom" column refers to the atom numbers in Fig. 1. "FC" is the function class, as listed in Table 1. "Dist." is the distance between an atom in molecule A and the nearest atom in molecule B. Note that no atom in molecule B is closer than 1Å to atom 6 of molecule A. Note also that although both atoms II and III of structure B are within 1Å of atom 1 of structure A, only the closer atom (#II) is taken for comparison. "$1.0\text{-}d^3$" is the distance weighting factor ($1Å^3 - \text{distance}^3$). $SM_{(B,A)}$ is the atomic similarity of two matched atoms as given in Table 2. In this case all matched atoms have the same functional class, except atoms 5, 7, and 9. The partial score is the distance weighting factor times the atom similarity, while the max score is a prefect spatial overlap and functional class match (11 x 1.0 x 32 = 352). The atom score is the partial score divided by 32. The 3D similarity of these two poses is the total of the atom scores = 0.699, with almost all of the dissimilarity being due to the unmatched atom 6 and the differing atom types (functional classes) of atoms 5, 7, and 9.

defined threshold. In the case of structures with the same number of similar structures, the sum of the similarity scores themselves is used to select the cluster center. All poses with a similarity score greater than the threshold are then placed in this cluster. The process is repeated, ignoring singletons and molecules already assigned to other clusters, until no molecule pairs are found within the user-defined similarity threshold (Fig. 2). The advantages of this clustering method over the popular and probably faster K-means clustering [14] is that the program determines both

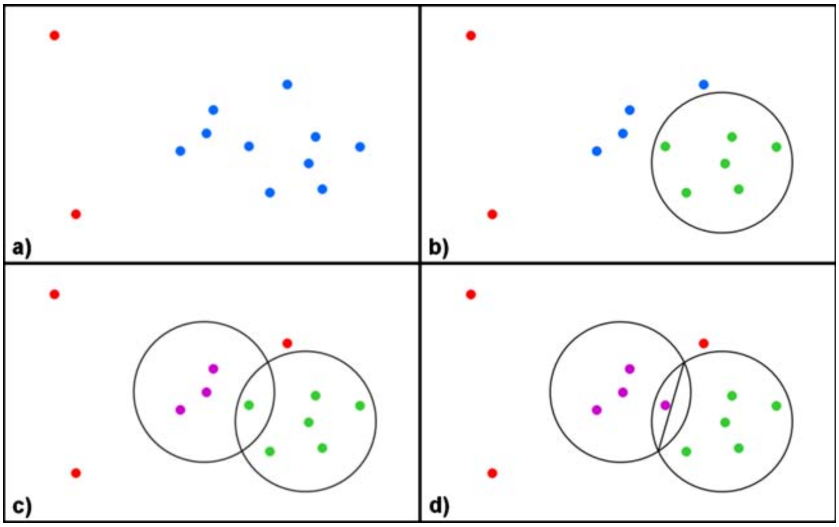


Fig. 2 Simple example of how clustering is carried out. a) A set of points to be clustered. The two red points are beyond the distance (similarity) threshold of all other points and therefore can be immediately be classified as "absolute" singletons. b) The first cluster center is selected as that point with the most other points within the distance threshold (circle). All points within the distance threshold (green points) are marked as members of this cluster and removed from further consideration. c) Second cluster center selected, which only has two other members (magenta points). At this stage there is still a single point left unclustered. Although it is within the distance threshold of other points, it is not within the distance threshold of any cluster center, therefore it is flagged as a singleton. d) Refinement step: one point originally assigned to the first (green) cluster is found to be actually closer to the second (magenta) cluster center and changes cluster membership

the number of clusters and the cluster centers based on the poses themselves, reducing the guesswork for the user and allows the program to always generate the same results given the same input.

Because cluster members are allocated sequentially, based on when a cluster center was identified, it may be that some members of a cluster actually fit better in a later defined cluster, i.e., are closer to another cluster center. For this reason, a final check for which cluster each pose best belongs to is carried out. During this stage some clusters may disappear because too many of their members are reassigned to other clusters. If only one member of a cluster remains at the end of the refinement step, it is relabeled a singleton. Because of the way cluster memberships are assigned, this is a rare, but not impossible event.

Finally, a copy of the input pose file is written out in which two new properties per pose are added: "cluster_number" and "cluster_similarity". Singletons are assigned to cluster number 99999 and given a cluster similarity of zero. The order of the poses in the file reflects their order in the input file, i.e., no sorting of the output poses based on cluster number or similarity is done.

Examining clusters by activity

It would be of considerable interest to see whether actives molecules tend to cluster together in 3D separate from inactive molecules. To examine this, the NCI set of compounds screened for anti-viral (HIV) activity (http://dtp.nci.nih.gov/docs/aids/aids_data.html, file aids_conc_may04.txt) was docked against HIV-1 protease. The collection consists of 43,850 compounds annotated as being either "confirmed active" (CA), "confirmed moderately active" (CM), or "confirmed inactive" (CI). Note that activity is against the whole virus and does not necessarily imply activity against HIV-1 protease. The compounds were filtered as before (Table 4), but with less stringent restrictions due to the small number of compounds: 800 > MW > 0, no unusual elements, <5 undefined chiral centers and <13 rotatable bonds. Altogether 34,082 structures (78%) were accepted. These were expanded for undefined chiral centers and alternative charge and tautomer states giving 129,435 total structures. The massive expansion was primarily due to accepting up to four undefined chiral centers, resulting in 16 enantiomers each. As before, these were docked to HIV-1 protease enforcing at least one H-bond to one of the catalytic aspartic acid residues. After removing low-scoring isomers, the top scoring 20,000 poses were clustered in 3D with a similarity threshold of 0.7.

The question to be answered is whether the various activity types tend to cluster together or do the 3D clusters contain molecules of mixed activity types. The analysis is complicated by the varying sized of the clusters, so all member of large clusters (>2 members) were broken down into all possible combinations of clusters of 2 molecules, e.g., a cluster of four molecules (A,B,C,D) would break down into six 2-molecule clusters: (AB), (AC), (AD), (BC), (BD), and (CD). Since there are three activity types: active (A), moderately active (M) or inactive (I), there are six possible combinations: (AA), (AM), (AI), (MM), (MI), and (II), whose predicted random distribution by activity type in 2-member clusters is easily calculated from the activity populations. Comparison with the observed distribution would then show whether molecules of the same activity tend to cluster together, cluster preferentially with other activity types ("cross-clustering") or show a random distribution.

Table 4 Filtering and expansion of the NCI compound collection

250,251	Starting SMILE strings
−12,966	Bad atoms (non- H, C, N, O, S, P, F, Cl, Br, I, Si)
−16,592	Duplicates
−1076	Molecular weight <150
−6690	Molecular weight >600
−21,622	> 2 undefined chiral centers
−20,213	> 8 freely rotatable bonds
161,402	Acceptable compounds (64.5%)
+59,213	Stereochemical expansion
+37,568	Alternate charge states (pKa 6–8)
−2376	Duplicates generated by expansion
255,807	Molecules to dock

Surprising was the number of duplicates (6.6%) and compounds containing elements not usually associated with modern drug design (5.2%). Essentially the entire periodic table was represented except for the noble gases, the transuranics, technetium, promethium, astatine and radium; with boron, arsenic, tin, and platinum each occurring in more than a thousand compounds.

Results

Test case

The National Cancer Institute's (NCI) compound collection (http://cactus.nci.nih.gov/ncidb2/download.html, August 2000, file NCI_aug00_SMI.sdz) was selected as a source of compounds. The 250'251 SMILES strings were processed to create a set of suitable compounds for docking. The collection contained an unexpectedly large number of compounds containing atoms generally not encountered in medicinal chemistry, as well as a large number of duplicates (Table 4). In addition, a Pipeline Pilot (SciTegic, San Diego, CA, USA) protocol was written to reject molecules on the basis of molecular weight, number of undefined chiral centers and number of freely rotatable bonds, which all together removed about one-third of the starting compounds (Table 4). Expanding the set for undefined chiral centers and alternative ionization states (pKa 6–8, as determined by the "enumerate ionization states" component of Pipeline Pilot) brought the number of structures to dock up to 255'807. No expansion of possible tautomeric states was done.

As protein target, HIV-1 protease was selected (liganded with 2,5-dibenzyloxy-3-hydroxy-hexanediotic acid *bis*-[(2-hydroxy-indan-1-yl)-amide]: PDB accession code 1D4I [15]). All solvent atoms were removed, including the one that often bridges the ligand and the flaps of the protein, and both catalytic residues (Asp-25 and Asp-125) were left in the deprotonated, anionic state. Docking was performed with Glide_SP (Schrödinger, LLC, Portland, OR, USA) with the constraint that at least one hydrogen-bond exist between the ligand and either of the catalytic aspartate residues. Only the single best pose per docked structure, as measured by its overall docking score "glide_gscore", was taken for further analysis.

The top 10,000 docked molecules (1 pose each) were selected and clustered in 3D as described above. 21 runs were made varying the similarity threshold from 0.0 to 1.0 in steps of 0.05 to examine how the number and size of the clusters generated, as well as the number of singletons, varied (Fig. 3). As expected, with a similarity threshold of 0.0, only a single cluster of all 10000 poses was created. At the other extreme, a similarity threshold of 1.0, no clusters were found, only 10,000 singletons, since the structures docked were all unique. In between, the number of clusters slowly increases to a maximum of almost 1350 at similarity threshold 0.55, before decreasing slowly down to zero at a similarity threshold of 1.0. Average cluster size (discounting singletons) decreases very rapidly to asymptotically approach 2, the minimum cluster size (Fig. 3b). The number of singletons increases slowly at first, but between 0.3 and 1.0 increases quite linearly (744 per 0.05 step with a correlation coefficient of 0.995).

To get a qualitative impression of what a given similarity threshold means with respect to chemical and spatial overlap, the 20,000 top poses were clustered with a similarity threshold of 0.5. Poses from a single very large cluster are shown in Fig. 4 as a function of decreasing similarity threshold. Similarity threshold 1.0 gives a single pose, that of the cluster center. Thresholds down to 0.85 perfectly preserve the scaffold with minor substitution changes (methyl, ethyl, halogen) and near-perfect spatial alignment. At a threshold of 0.8 there is the first scaffold change (dibenzofuran for biphenyl), although still with good spatial alignment. Also at this point for the first time the carbamate between the phenyl and biphenyl groups is replaced (with a carbamide). At 0.75 more chemical variation is seen, although spatial alignment is still very good. From 0.7 to 0.6 both chemical similarity and spatial alignment continue to deteriorate. From this example and others, it is recommended to select a similarity threshold somewhere between 0.7–0.8. Lower thresholds give few, larger clusters and therefore can reduce the list of hits more, but the pose selected to represent the cluster may no longer be a good representative of all members of the cluster. Higher thresholds give excellent alignment, but result in many very small clusters and many singletons and therefore little reduction in the hit list.

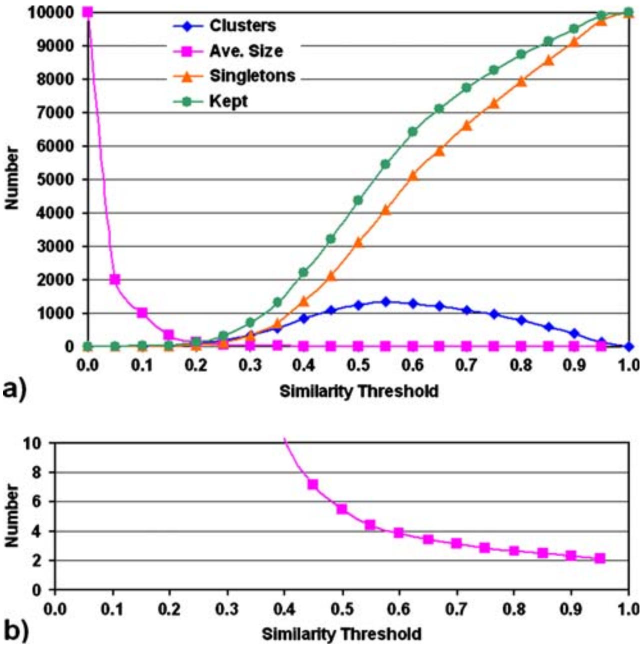


Fig. 3 a) plot of the number of clusters (blue diamonds), singletons (orange triangles), average cluster size (magenta squares) and number of clusters + singletons (green circles) as a function of similarity threshold for 10,000 top-docked NCI compounds docked against HIV-1 protease. At very low similarity thresholds there are very few, very large clusters. The number of clusters reaches a maximum at thresholds 0.50–0.60 before decreasing again to very few, but now very small, clusters as one approaches a threshold of 1.0. The number of singletons also starts off very small at very low similarity thresholds, but begins to increase rapidly at around 0.3 until only singletons exist at a threshold of 1.0, since the docking set did not contain any duplicate molecules. b) is a much expanded view of the average cluster size, showing how it asymptotically approaches 2.0 with increasing similarity threshold

Analysis of 3D clusters vs. activity

To check whether compounds that cluster together tend to have similar activity, the clusters of the NCI anti-viral screened molecule set docked to HIV-1 protease were analyzed. Docking of the 129,435 expanded structures generated 88,499 acceptable poses. Accepting only the single top-scoring pose per molecule gave 22,538 unique molecular poses of which the top 20,000 were clustered in 3D using a similarity threshold of 0.7. 1789 3D clusters

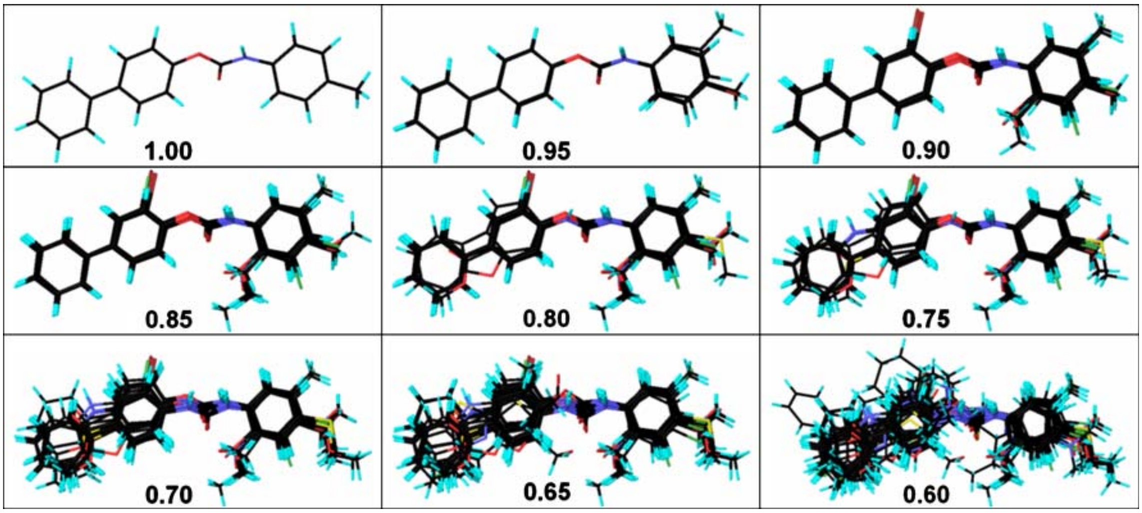


Fig. 4 Member poses of a large cluster of NCI molecules docked to HIV-1 protease overlaid as a function of similarity threshold. 1.00 indicates the molecule at the cluster center. Note that the cluster center scaffold is rigidly maintained down to a similarity of 0.80, with only relatively minor substituent changes and very good spatial overlap. With decreasing similarity threshold, both the chemical variability and spatial overlap of the molecules gradually spread

were generated with sizes ranging between 2 and 17 members and containing a total of 4986 poses (24.9%). The rest of the poses were singletons. The clustered molecules consisted of 52 actives (0.71%), 123 moderately actives (2.42%) and 4811 inactives (96.87%). Reducing large clusters (>2 members) to all possible 2-member clusters to simplify the analysis generated 6603 2-member clusters with the following activity distribution: 221 actives (1.67%), 301 moderately active (2.28%) and 12,684 inactives (96.05%). Note that the activity distribution has changed because of the bias generated by large clusters. A 2-member cluster generates only a single 2-member cluster, whereas a 17-member cluster generates 136 2-member clusters $(N * (N - 1)/2)$.

Table 5 show the expected distribution of the 2-member cluster activity types versus the distribution actually observed. It is clear that molecules of the same activity class tend to cluster together, the active compounds being almost 43× more likely to cluster with other actives than expected from a random distribution. Note that although the 1.02× enrichment in (II) clusters does not seem significant, because the (II) clusters already represent 94% of all the clusters, even if all inactives became (II) clusters, this would represent only a 1.04x increase. It is also interesting that there is a significant increase in observed “cross-clustering” between active and moderately active molecules and that “cross clustering” between active and inactive molecules is more reduced than between moderately active and inactive ones. These observations show that, at least in this case, there is a strong tendency of molecules to 3D cluster by activity class. This is even more surprising given the fact that anti-viral activity of these molecules is defined

Table 5 Analysis of 3D clusters by activity

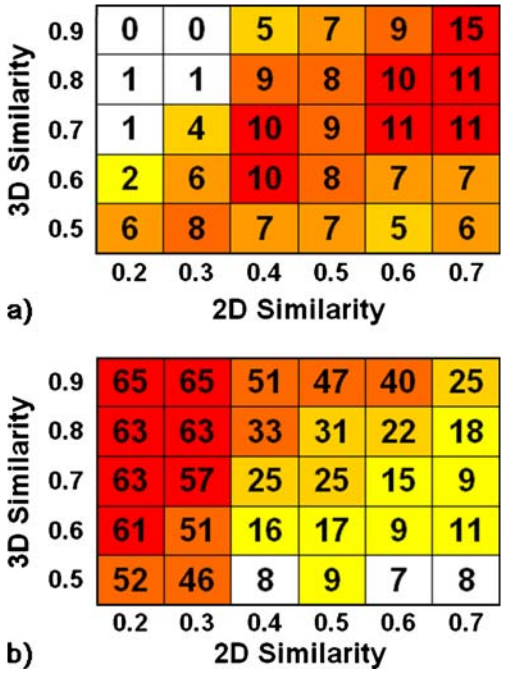
Theoretical	A (.0167)	M (.0228)	I (.9605)
A (.0167)	0.000279	0.000381	0.016040
M (.0228)	0.000381	0.000520	0.021899
I (.9605)	0.016040	0.021899	0.922560
Expected	A	M	I
A	1.8	2.5	105.9
M	2.5	3.4	144.6
I	105.9	144.6	6091.7
Observed	A	M	I
A	79	9.5	22
M	9.5	28	113
I	22	113	6207
Enrichment	A	M	I
A	42.90	3.78	0.21
M	3.78	8.16	0.78
I	0.21	0.78	1.02

The top table shows the proportion of the various activity combinations for 2-member clusters for the three activity types A=active, M=moderately active, I=Inactive, if they were perfectly evenly distributed by the frequencies of their occurrence (given in the headers). The second table shows how the 6603 2-member clusters found should be distributed based on a perfectly even distribution of the activity classes. The third table shows the actual observed distribution of 6603 2-member clusters generated by docking followed by 3D clustering. Note that each off-diagonal element is actually one-half the observed number of cluster since each cross-cluster appears twice in the matrix, i.e., cluster AM = cluster MA. The final table shows the enrichment of 2-cluster types by dividing the elements of the observed distribution (third table) by those of the expected distribution (second table).

against the whole HIV-1 virus and not just the protease under examination here.

Comparison of 2D with 3D clustering

To examine the effects of 3D clustering poses as a function of their 2D similarity, 6 sets of high-scoring NCI compounds docked to HIV-1 protease were generated based on their 2D similarity. Using the functional class fingerprint, 4 atom radius (FCFP_4) of Pipeline Pilot, compound sets were generated with similarity thresholds of 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7. Each set had 65 compounds, this limit being determined by the maximum number of compounds in the most restrictive set (0.7 similarity). For the larger sets, compounds were randomly selected to give exactly 65 members each. Each of these sets were then clustered in 3D varying the 3D similarity thresholds from 0.5 to 0.9 in steps of 0.1 and observing the number of 3D clusters formed and singletons generated. These sets were then plotted by the number of clusters and singletons found as a function of their 2D and 3D similarity thresholds (Fig. 5). At the most stringent 3D similarity threshold (0.9), only with a molecule set with 2D similarity of 0.4 or higher are any clusters found. Even at 0.7 2D similarity, over a third of the molecules are singletons when clustered with a 3D similarity threshold of 0.9. At the other extreme, the number of clusters generated with a 3D similarity threshold of 0.5 does not vary much as a function of 2D similarity, although the size of the clusters, and concurrently the number of singletons, changes dramatically in going from a 2D similarity threshold of 0.3 (average cluster size 2.4, 46 singletons) to 0.4 (average cluster size 8.1, 8 singletons). It is clear from this analysis that high 2D similarity is required, but not necessarily sufficient for high 3D similarity: a similar binding mode (spatial overlap) is also required.



3D Similarity	0.2	0.3	0.4	0.5	0.6	0.7
0.9	0	0	5	7	9	15
0.8	1	1	9	8	10	11
0.7	1	4	10	9	11	11
0.6	2	6	10	8	7	7
0.5	6	8	7	7	5	6

3D Similarity	0.2	0.3	0.4	0.5	0.6	0.7
0.9	65	65	51	47	40	25
0.8	63	63	33	31	22	18
0.7	63	57	25	25	15	9
0.6	61	51	16	17	9	11
0.5	52	46	8	9	7	8

Fig. 5 Number of a) clusters and b) singletons as a function of 3D similarity threshold for 6 sets of 65 molecules generated by varying 2D similarity thresholds from 0.2 to 0.7. The number of clusters and singletons are colored white-yelloworange-red with increasing size for clarity. An obvious increase in the number of 3D clusters and consequent decrease in singletons are seen for the sets of compounds with increasing 2D similarity with the largest jump being between compounds generated with a 2D similarity of 0.3 and those generated with a 2D similarity of 0.4

Performance and limitations

All execution time tests were performed on a Hewlett-Packard xw-8000 Linux workstation with 2 Intel Xeon CPUs (3.2 GHz, 1 MB cache each) and 1 GB RAM. The program is not parallelized, but two separate runs of the program can be carried out simultaneously on the workstation without performance degradation. The program performs two basic operations: calculation of an N × N similarity matrix, where N is the number of poses to be clustered, and clustering the poses based on these similarities and a user-supplied similarity threshold. The speed of calculating of the similarity matrix was increased by ignoring hydrogen atoms (although using their presence or absence to define the function class of the parent heavy atom) and only calculating the similarity of the larger against the smaller of each pose pair. The clustering algorithm works by sequentially finding cluster centers with the most poses within the similarity threshold. Each pose is scored according to the number of other poses with a similarity score greater than the threshold and the similarity score itself, ignoring poses already assigned to previously defined clusters. This was originally rescored after each new center was located, but proved to be quite slow. Re-writing the algorithm to ignore known singletons altogether and to calculate this score just once, but then subtract the contributions from newly discovered cluster members as they are allocated, produced an enormous speed-up, especially for very large pose sets (up to 100×). Figure 6 shows the execution time for the two steps, similarity calculation and clustering, varying with the number of poses to be clustered. Care was taken to ensure that the average molecular size was similar in each set (22.5–22.8 non-hydrogen atoms). Both processes increase roughly as the square of the number of poses to be clustered, although the time needed to generate the similarity matrix is roughly 45× longer than for clustering. For this reason it was made possible to export the similarity matrix as an ASCII text file so that re-running the program for the same pose set, e.g., with a different similarity

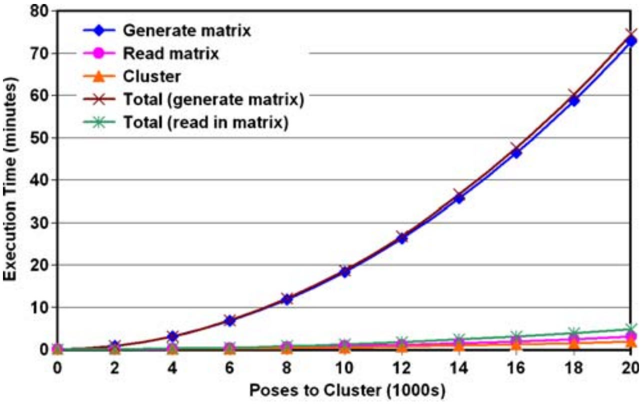


Fig. 6 Execution times as a function of the number of poses to be clustered: similarity matrix generation (blue diamonds), clustering (orange triangles) and total execution time (brown X's). Also shown are the times when the similarity matrix from a previous run is read in, instead of being recalculated (magenta circles) and the total execution time in this case (green stars)

threshold, and reading in the similarity matrix rather than re-calculating it would be significantly faster. In fact, this reduced the overall execution time 15-fold.

Because cluster membership is assigned sequentially as highly populated cluster centers are located, it may be that a pose actually fits better in a later defined cluster. To check for this, a final refinement step is performed to assign each pose to that cluster where its similarity is highest to the cluster center. Figure 7 show how the cluster sizes are changed in a case where 2000 poses were clustered with a similarity threshold of 0.4 before and after refinement. In this particular case 22.0% of the poses changed their cluster membership, resulting in an improvement in the average similarity (excluding cluster centers) from 0.4749 to 0.4908. These effects are far less dramatic at higher similarity thresholds where the “membership criteria” are more stringent to begin with. In the same case above, but with a similarity threshold of 0.8, only 1.5% of the poses changed cluster during refinement with no significant increase in the average similarity.

The program can currently only read molecular poses in SDF format. Hydrogen atoms are ignored. There is a limit of 20,000 poses that can be clustered. This is because the entire similarity matrix (up to 4×10^8 elements) is held in RAM during execution. Each pose can have no more that 200 atoms, including hydrogen atoms. This corresponds to molecules of roughly 1600 Dalton and should cover almost all compounds likely to be encountered in drug discovery. Only molecules composed of C, N, O, S, P, F, Cl, Br, I and Si have atoms with defined functional classes. Hydrogen atoms are accepted, but ignored, except in defining the functional class of their parent heavy atom. If the protein to which the molecules were docked is read in for weighting the individual atoms according to their interactions with the protein, this must be in protein data base (PDB) format. Up to 10,000 non-hydrogen protein atoms (ca. 1250 amino acid residues) can be read in.

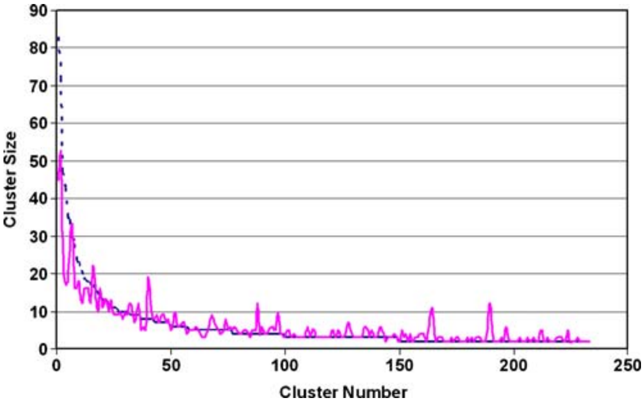


Fig. 7 Cluster size before (dotted blue line) and after (solid magenta line) refinement for 2000 poses with a similarity threshold of 0.4. Since cluster centers are assigned according to the number of other molecules that would cluster with it, it is not surprising that before refinement the first cluster is the largest, steadily decreasing to the minimum cluster size of two. The refinement step assigns members to the closest cluster center, not the one that first picked it up, essentially taking from the rich and giving to the poor, but only if the poor (small) cluster center is actually more similar to the molecules in question. In this particular case, it can be seen that the first cluster has lost almost half of it members to smaller clusters, while two minimal clusters (clusters 164 and 190) grew to over ten members. Note that for real-life cases the clustering threshold would be higher (usually 0.7–0.8) and the number of molecules changing cluster would be much smaller

Discussion

A FORTRAN program has been described that allows the clustering of up to 20,000 docking poses based on both their chemical similarity and 3D spatial overlap. This should be superior to clustering based on 2D chemical descriptors alone, since in addition to these, one is also including information about the predicted binding mode. This clustering can be used to reduce a list of top scoring docking poses by selecting a single or few representative compounds from each cluster for biological testing. It is possible to weigh the contribution of the individual atoms of a pose based on the number of interactions it is making with the target protein, i.e., by its contribution to binding. Tests have shown that 3D similarity thresholds between 0.7 and 0.8 give poses that overlap well visually, yet produce clusters that are large enough to give a significant reduction in the hit list. Calculation of the similarity matrix is quite slow compared to the time required for the cluster poses based on this matrix, however, once the matrix has been calculated it is possible to export it so that for future runs, e.g., with different similarity thresholds, it only needs to be imported. This is about 15× faster altogether than re-calculating it. Within the limitations of the program (≤ 20,000 poses, each with ≤ 200 atoms, no "weird" atoms) the program is very robust.

Acknowledgements The author would like to thank Dr. Woody Sherman (Schrödinger, LLC) for suggesting weighting of the individual atoms' contribution to the overall similarity of the poses based on interactions with the protein and to thank Dr. Bernhard Rohde (Novartis Institute of Biomedical Research) for useful discussion and comments.

References

1. Willett P, Winterman V, Bawden D (1986) J Chem Inf Comput Sci 26:109–118 doi:10.1021/ci00051a005
2. Shemetulskis NE, Weininger D, Blankley CJ, Yang JJ, Humblet C (1996) J Chem Inf Comput Sci 34:862–871 doi:10.1021/ci950169+
3. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) J Chem Inf Comput Sci 41:702–712 doi:10.1021/ci000150t
4. Raymond JW, Willet P (2002) J Comput Aided Mol Des 16:521–533 doi:10.1023/A:1021271615909
5. Hattori M, Okuno Y, Goto S, Kanehisa M (2003) J Am Chem Soc 125:11853–11865 doi:10.1021/ja036030u
6. Raymond JW, Blankley CJ, Willett P (2003) J Mol Graph Model 21:421–433 doi:10.1016/S1093-3263(02)00188-2
7. Li W (2006) J Chem Inf Model 46:1919–1923 doi:10.1021/ci0600859
8. Mattos C, Rasmussen B, Ding X, Petsko GA, Ringe D (1994) Nat Struct Biol 1:55–58 doi:10.1038/nsb0194-55
9. Wu N, Pai EF (2002) J Biol Chem 277:28080–28087 doi:10.1074/jbc.M202362200
10. Cody V, Luft JR, Pangborn W, Gangjee A, Gueener SF (2004) Acta Crystallogr D Biol Crystallogr 60:646–655 doi:10.1107/S0907444904002094
11. Deng Z, Chuaqui C, Singh J (2004) J Med Chem 47:337–344 doi:10.1021/jm030331x
12. Schneider G, Neidhart W, Giller T, Schmidt G (1999) Angew Chem Int Ed 38:2894–2896 doi:10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F
13. Heyer LJ, Kruglyak S, Yooseph S (1999) Genome Res 9:1106–1115 doi:10.1101/gr.9.11.1106
14. MacQueen JB (1967) Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281–297
15. Andersson HO, Fridborg K, Löwgren S, Alterman M, Mühlman A, Björsne M et al. (2003) Eur J Biochem 270:1746–1758 doi:10.1046/j.1432-1033.2003.03533.x