molecular informatics

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Structuring Chemical Space: Similarity-based Characterization of the PubChem Database

Giovanni Cincilla,[a,b] Michael Thormann[c] and Miquel Pons*[a,b]

Abstract: The ensemble of conceivable molecules is referred to as the Chemical Space. In this article we describe a hierarchical version of the Affinity Propagation (AP) clustering algorithm and apply it to analyze the LINGO-based similarity matrix of a 500,000-molecule subset of the PubChem database, which contains more than 19 million compounds. The combination of two highly efficient methods, namely the AP clustering algorithm and LINGO-based molecular similarity calculations, allows the unbiased analysis of large databases. Hierarchical clustering generates a numerical diagonalization of the similarity matrix. The target-independent, intrinsic structure of the database , derived without any previous information on the physical or biological properties of the compounds, maps together molecules experimentally shown to bind the same biological target or to have similar physical properties

Keywords: LINGO, Molecular Similarity, Affinity Propagation Clustering, PubChem, Eigenmolecules

1 Introduction

The number of compounds recorded in chemical databases is growing rapidly. The NIH-sponsored public repository of molecular information, the PubChem database (http://pubchem.ncbi.nlm.nih.gov), exceeded 19 million compound records in September 2008, only 4 years after its implementation. This vast amount of information generates new opportunities for exploiting structure activity relationships in pharmaceutical research. In this scenario, the notion of chemical space and its characterization has become a central issue.

Underlying the practical applications of the chemical space concept is the Similar Property Principle, for which molecules that are structurally similar are likely to have similar physicochemical properties, [1,2] as well as similar interactions with targets (Active Analog Principle). [3-7] Chemical similarity-based clustering methods attempt to organize the chemical space by grouping molecules with similar properties. Complete reviews of clustering methods applied to chemical problems are available. [8] Clustering methods can be classified as hierarchical or non-hierarchical. The former generate classifications that are typically represented as dendrograms. In contrast, the latter generate a partition of elements. However, iterative clustering of exemplars, the representatives of each cluster, can be used to convert the method from non-hierarchical to hierarchical.

Very recently, Frey and Dueck introduced a new affinity propagation (AP) clustering algorithm. [9] This extremely efficient algorithm takes pairwise distances between data points as input in order to select those that form the centre of clusters. Thus the sum of the squared distances between data points in the cluster and their centres are minimal. In contrast to other clustering techniques, all points are initially considered potential exemplars and real-valued messages are recursively passed between points to indicate the “responsibility” of each data point to choose another data point as its exemplar and the “availability” of each data point to become an exemplar, on the basis of the accumulated “support” gathered from other points. Messages are updated at each iteration using simple formulas that search for the minima of an appropriate energy function. The computational cost increases linearly with the number of similarity values used. Thus this approach is well suited for the clustering of very large datasets with a relatively small number of significant similarities between data pairs. AP clustering can be applied to data in discontinuous non-metric space. This is the case of chemical space and pairwise distances derived from Molecular Similarities. A large number of molecular similarity measures have been described. [3, 4, 10-13] Most approaches use graph-based methods to identify molecules and to compute their similarities. We recently introduced LINGO pure text- based methods to compute molecular similarities and molecular properties. [14, 15] LINGO methods compare similarities directly from standardized SMILES or IUPAC names. [16] In contrast to fingerprint-based methods, direct comparison of SMILES or IUPAC names obviates the expansion of the associated molecular structures. In spite of their simplicity, LINGO methods have been shown to give

[a] *Institute for Research in Biomedicine Parc Científic de Barcelona. Baldiri Reixac, 10. 08028-Barcelona. Spain*

[b] *Departament de Química Orgànica. Universitat de Barcelona. Martí i Franquès, 1-11. 08028-Barcelona. Spain*

mpons@ub.edu Tel +34 934034683 Fax +34 934034683

[c] *Origenis GmbH, Am Klopferspitz 19a. 82152 Martinsried. Germany:*

Supporting Information for this article is available on the WWW under www.molinf.com comparable results to widely used structure-derived fingerprint methods to retrieve active compounds for a variety of activity classes from a random set. [17]

Here we used AP clustering, together with LINGO-based similarity calculations, to analyze the structure of PubChem. Using a subset of 500,000 molecules, we derived a stable hierarchical clustering structure. We show that molecules with similar physical properties or that are experimentally known to bind to the same target are clustered together.

The use of molecular basis sets to represent chemical space was previous proposed by Oprea and Gottfries in their ChemGPS method. . [18] In that procedure, external “satellite” molecular vectors with components defined by a number of molecular descriptors provide the positions of molecules in the drug-like chemical space Also, the SIBAR method uses the similarity of molecules to a set of reference compounds to estimate molecular properties. [19] The chemical space has analogies with a high dimension vectorial space. Recently, Raghavendra and Maggiora used a generalized Fourier analysis to describe chemical space using basis sets of orthogonal abstract molecular vectors associated with chemical species.[20] Hierarchical clustering provides a method for the diagonalization of the similarity matrix and identifies sets of “eigenmolecules” that are mutually dissimilar (“orthogonal”) and related to most of the molecules in the database.

2 Methods

PubChem database processing

The PubChem database was downloaded on 03/05/2008 from the PubChem server. All entries were converted into canonical SMILES using JChem software. [21] The database was filtered to remove the following: compounds containing elements other than C, N, S, O, H, P and halides; double entries; SMILES shorter than 4 characters; and molecules with more than 9 rings or with a molecular weight > 600D. The processed database contained 16,021,418 entries. Hierarchical clustering was applied to a random subset of 500,000 molecules.

Similarity calculations

LINGOsim [14] pairwise molecular similarity was calculated from standardized SMILES-derived 4-character LINGO profiles using equation (1)

$$S(A,B) = \frac{\sum_{i=1}^{l} 1 - \frac{|N_{A,i} - N_{B,i}|}{|N_{A,i} + N_{B,i}|}}{l} \qquad (1)$$

where $N_{A,i}$ is the number of LINGOs of type *i* in molecule A, $N_{B,i}$ is the number of LINGOs of type *i* in B, and *l* is the number of LINGOs contained in either molecule A or B.

Affinity propagation algorithm clustering.

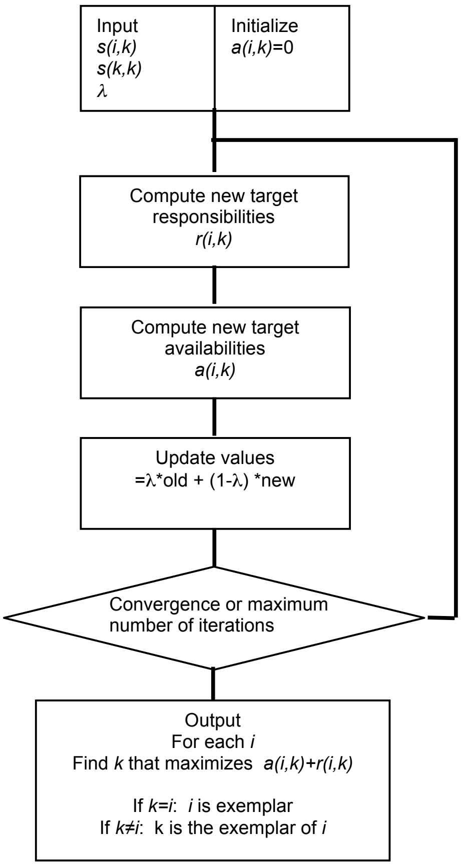
AP clustering takes a list of similarities *s(i,j)* and preferences *s(k,k)* as input. The preference input value indicates the *a priori* likelihood that molecule *k* becomes an exemplar. This value controls the average number of clusters formed. The use of the estimated median value of similarities in the database (0.0533), as recommended in the original publication, leads to a number of exemplars of around one tenth of the points to be clustered. The Matlab implementation of the AP clustering algorithm was downloaded from http://www.psi.toronto.edu/affinitypropagation/ (accessed on 03/06/08). This implementation uses negative distances as input, where distance is set to 1- *S(i,j)*. Briefly, AP clustering is based on the recursive updating of two types of real-valued messages: the responsibility of point *i* towards candidate exemplar *k*, *r(i,k)* and the availability of candidate *k* to be the exemplar of *i*, *a(i,k)*. Availabilities are initialized to zero. Updates are carried out using the following rules.

$$r(i,k) \leftarrow s(i,k) - \max\{a(i,k') + s(i,k')\} \qquad (2)$$

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\} \qquad (3)$$

$$a(k,k) \leftarrow \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\} \qquad (4)$$

The AP clustering algorithm pseudocode is shown in Scheme 1.


Input
s(i,k)
s(k,k)
λ
Initialize
a(i,k)=0
Compute new target responsibilities
r(i,k)
Compute new target availabilities
a(i,k)
Update values
=λ*old + (1-λ) *new
Convergence or maximum number of iterations
Output
For each i
Find k that maximizes a(i,k)+r(i,k)
If k=i: i is exemplar
If k≠i: k is the exemplar of i

Scheme 1. Pseudocode for AP-clustering

Iterative clustering.

AP clustering was carried out starting from a random ordered database by clustering non-overlapping sequential sets of 3,000 molecules. The size of the subsets was arbitrary and was chosen for computating time reasons. The resulting exemplars, which we call “parents”, were submitted to a new clustering procedure in groups of 3,000. This approach produced a new, smaller set of exemplars, which we refer to as “grandparents”. The procedure was repeated to generate fewer than 3,000 “great grandparents”, which were clustered to generate the highest level exemplars, which we call “ancestors”. The composition of the starting subsets and the clusters derived depends on the original ordering of the database. However, this procedure generates a hierarchy for each molecule in the database, which is used to reorder the original dataset so that molecules sharing common exemplars at distinct clustering levels are placed together, with priority given to sharing exemplars that belong to the lower levels of the clustering hierarchy. The procedure is illustrated by an example in the supplementary information. Clustering is refined at each iteration both globally (by bringing together molecules that were located in distant positions in previous database orderings and therefore not included in the same subset) but also locally (as the molecules that are compared within the same subset become more similar, clustering generates a finer discrimination between similar compound classes). The complete procedure (generation and clustering of non-overlapping sets of 3,000 molecules to produce a new set of “parents” and, subsequently, “grandparents”, “great-grandparents” and “ancestors”) was repeated until a stable clustering of the complete dataset was achieved. Convergence was monitored through the average similarity of sets of adjacent molecules in the ordered database and by comparing the cluster composition in different iterations using ClustSim. The average time for a process iteration is about 12 hours in a single PC running with a Intel Core 2 quad processor on a GNU/Linux platform.

ClustSim values are defined as follows: for a molecule α that belongs to cluster A in one round and to cluster B in a second run:

$$ClustSim_{\alpha}^{A-B} = \frac{c-1}{a+b-c} \quad (0 \leq ClustSim < 1) \qquad (5)$$

where: *a* is the number of molecules in cluster A; *b* is the number of molecules in cluster B; and *c* is the number of molecules present in both cluster A and cluster B. This coefficient is zero when α is the only molecule shared by clusters A and B and it increases with the number of common molecules in both clusters. The maximum value of ClustSim is *(k-1)/k* for identical clusters of size *k*. For large clusters, this value approaches one, although for small clusters the contribution is lower. For singletons the contribution becomes zero. The similarity between clusterings of the same set of molecules is the sum of ClustSim for all molecules in the set. The maximum value was estimated by computing the similarity between two identical cluster sets with the same distribution of cluster sizes.

Incompleteness error.

Following Raghavendra and Maggiora, [20] a limited set of molecules can be used as a basis set to describe a much larger database. The error associated with the incompleteness of a basis set of p molecules can be quantified by the sum of the components of molecules in a test set that are not described by this basis set using equation (6).

$$\varepsilon^2 (p) = 1 - \mathrm{Tr}(S^{T}\mathbf{S}^{-1}S) \qquad (6)$$

Where *S* is the *p* x *n* dimensional matrix formed by the similarities between the *p* molecules forming the chosen basis set and the *n* molecules of a test set. **S** is the *p* x *p* similarity matrix containing the similarities between the members of the basis set. The error decreases with the size of the basis set but it also depends on the composition of the set and, in particular, on its capacity to capture the diversity of the complete database. The larger the diversity captured by a basis set of a given size, the lower the incompleteness error.

Biological activity classes.

To test whether the LINGOsim-AP-derived clusters mapped together compounds with known activity for the same targets, we used 2,950 ligands for 40 distinct targets taken from the Directory of Useful Decoys (DUD). [22] For each ligand, the DUD database contains in addition 36 physically similar, but inactive decoy molecules. Thus the DUD database is a very strict benchmark for virtual screening. The total number of molecules is 98,266. Details of the database are given as supplementary material. AP clustering was performed on the DUD database and on a database generated by adding the 2,950 DUD ligands to the original 500,000-molecule database.

The distribution of the ligands for each target in the different clusters was compared with the distribution of an equivalent number of decoys of the same target or random molecules selected from the complete databases. In order to compare the results of different targets, which differ in the number of ligands, a cluster specificity index (CSI) was defined as

$$CSI(i) = \frac{N^i_{\max}}{N^i_C} \qquad (7)$$

where N^i_C is the number of clusters that contain at least one ligand of target *i* and $N^i_{\max}$ is the maximum number of clusters that can be occupied, which is the number of ligands for this class. Thus, the CSI value is numerically identical to the average number of ligands in the active clusters.

3 Results

Iterative hierarchical AP clustering.

For a given database, the similarity matrix contains all the pairwise similarities between the elements. The intrinsic structure of the database, as captured by a given combination of molecular descriptor(s) and similarity metric, is contained in the similarity matrix and it can be used to organize the database in two ways:

i) Elements can be grouped in clusters showing similarity within the cluster that is significantly higher than with members of other clusters.

ii) Individual elements of the database can be ordered in a list with the property that pairs of elements that are closer in the list are more similar than more distant pairs. The associated similarity matrix would have maximum values close to the diagonal.

Clustering techniques are either hierarchical or non-hierarchical. Hierarchical methods provide dendrograms (tree diagrams) that allow the definition of clusters at different levels of similarity. Elements high in the hierarchy will connect a large number of elements with low similarity. Lower hierarchy elements nucleate smaller clusters with higher similarity.

The original AP algorithm is non-hierarchical. Compared to other clustering algorithms, it has the advantage that it is extensively data-driven and does not require the predefinition of either the number of clusters or the cluster centers. In addition, this algorithm is robust and efficient and can be used with large datasets and discontinuous non-metric spaces. Each AP round generates a set of exemplars that are the centers of the clusters into which the complete database has been divided. AP clustering can be used iteratively to cluster exemplars, thereby resulting in hierarchical clusters. Here we performed four-level clustering (parents, grandparents, great grandparents and ancestors) to link the initial 500,000 molecules to ca. 100 ancestors, the top level exemplars. Figure 1 shows a representative set of structures from a low level cluster and their higher order representatives.

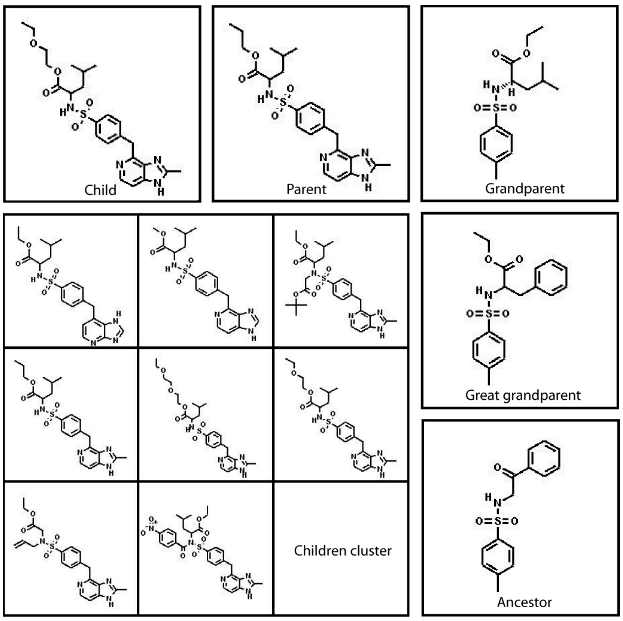


Figure 1. Representative example of a LINGOsim-based AP cluster and its associated higher level exemplars.

AP clustering cannot be applied directly to the complete set of 500,000 elements. The list of molecules in the database (in an initially arbitrary order) was divided into smaller subsets of adjacent molecules, which were hierarchically clustered. At the higher levels of the hierarchy, the exemplars from distinct subsets can be combined and clustered. As a result, elements from different subsets may have common ancestors. This initial hierarchical clustering allows the reordering of the molecules in the database, such that molecules with common ancestors are grouped. The complete procedure (division in small subsets of adjacent molecules, hierarchical clustering, and reordering of the database) is repeated until a stable clustering arrangement is obtained.

Clustering convergence and cluster structure

Each cycle can generate a potentially distinct set of clusters. The similarity between the composition of the clusters is given by the integral of ClustSim (see methods section), which compares the elements that are present in the clusters to which a given molecule belongs in two different clustering arrangements and adds the results for all the molecules in the database. Figure 2 shows the evolution of the similarity between clusters until convergence is reached after about 25 cycles. The similarity value at convergence is 94.8% of the value obtained by comparing identical clusters. This value corresponds to about 2.5% of all molecules being in a different cluster to that of the previous cycle.

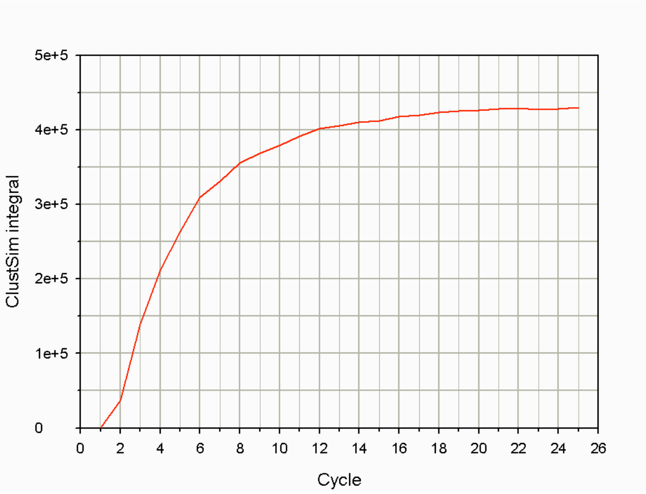


Figure 2. Convergence of cluster composition in successive hierarchical AP-clustering of a random set of 500,000 molecules from PubChem.

Convergence can also be estimated by comparing the average similarity within non-overlapping consecutive sets of 3,000 molecules in the ordered database (level 1) and sets of 3,000 exemplars of the higher levels (Figure 3). The average similarity of sets measured at the higher levels tends to decrease with respect to a random set of similar size. The average pairwise similarity between the final 101 ancestors of the database was only 0.02. This value was expected as the clustering procedure tends to select the most dissimilar molecules at the highest clustering level. Iterative hierarchical coupling provides a numerical approach to the diagonalization of the similarity matrix. Figure 4 shows the average similarity between molecules as a function of the distance of their positions in the ordered database after the process has converged.

Bioisoster molecules are related by the exchange of broadly similar atoms or groups, thereby leading to similar biological properties in a given context. The BIOSTER database provides a list of pairs of bioisoster molecules extracted from the literature in a range of contexts, including drugs, agrochemicals, enzyme-inhibitors and pro-drugs. Previous results showed that LINGO-derived similarities can statistically identify biosioster pairs: 95% of non-bioisoster molecules have LINGOsim values lower than 0.17. Thus higher similarity values suggest similar properties. [14] Using this threshold, bioisoster molecules are expected to be located in positions ±60 from a given compound in the ordered set.

The LINGOsim average inside clusters of different levels is shown in Figure 5. The arrows indicate the 0.17 similarity threshold of bioisoster molecules. Clusters generated by AP and LINGOsim group molecules with average pairwise similarities well above the bioisoster threshold. The average similarities at the second and third levels are still significant, thereby suggesting that physical properties and activity data from molecules sharing a “great grandparent” are statistically correlated.

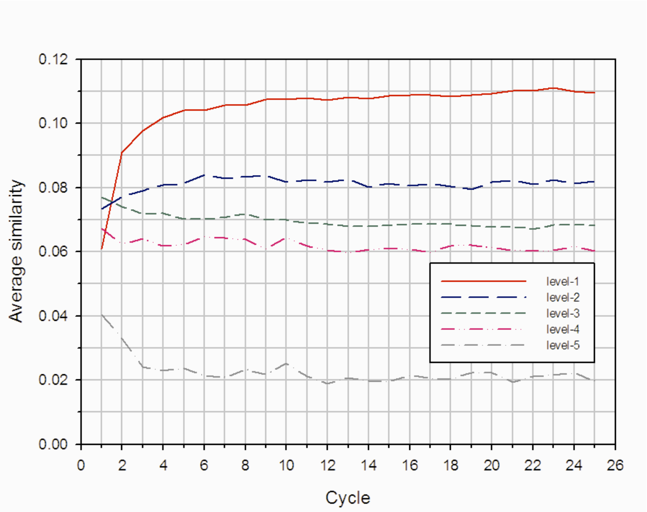


Figure 3. Average similarity within non-overlapping sets of adjacent molecules at different iterations during AP-clustering. The sets contained 3,000 molecules, except for the higher clustering levels where the total number of exemplars was used.

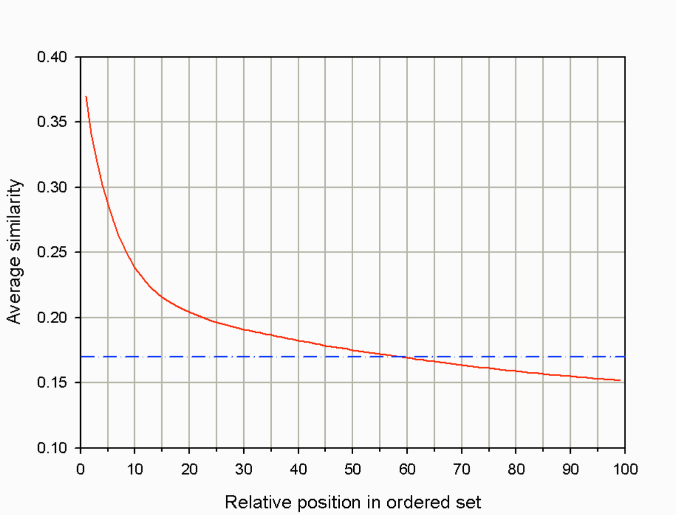


Figure 4. Average pairwise LINGOsim values of as a function of their relative position in the converged ordered set. The 0.17 threshold for bioisoster molecules is indicated.

Cluster stability.

To check the stability of the clustering algorithm with respect to the initial random ordering of the database, we performed a second complete clustering process, starting from a different random order of molecules. We then compared the similarity of the low level exemplars (parents) obtained in both cases. The results are shown in Figure 6a.

161,662 molecules (corresponding to 32.3% of the whole database) had exactly the same parent assigned in both clustering experiments. For the remaining 67.7% of molecules, the parents assigned in the two clustering experiments had varying degrees of similarities.

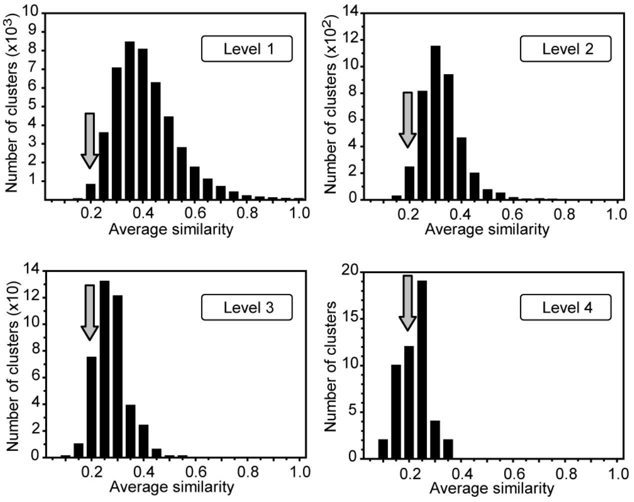


Figure 5. Intracluster average pairwise similarity. The number of clusters with an average pairwise similarity between S and S-0.05 is represented as a function of the similarity value S. Only clusters containing more than one element are considered. The members of a level i (i > 1) cluster are level i-1 exemplars. The arrows indicate the 0.17 threshold for bioisosters.

Figure 6b compares the distribution of similarities of 338,338 pairs of parents assigned to the same molecule in different clustering experiments with that of 7,161 bioisoster pairs from the BIOSTER database and the same number of random pairs of molecules. The similarity distribution of pairs of parents resembles that found in the BIOSTER database and clearly differs from the distribution expected for random pairs. These results show that the clusters generated by the method described here, starting from different database orders, are equivalent from the point of view of their structural and biological relevance.

However, the exact composition of the clusters is only partially maintained. The ClustSim coefficient between the two cluster sets is 19.9% of the expected value for identical clusterings. This value corresponds to the exchange of ca. 47% of the molecules between clusters. Equally acceptable (degenerate) clustering arrangements are possible. The similarity distributions show that degenerate clustering arrangements select exemplars that have similarities typical of bioisosters. This observation thus suggests that degenerate clusters have similar chemical and biological relevance.

Clustering stability is a crucial issue defined as the capacity of the algorithm to consistently generate the same clustering from the same dataset. Here we show that AP-LINGOsim clustering generates a stable clustering, from the point of view of bioisosterism. Thus, the resulting clustering captures functionally relevant aspects of the intrinsic structure of the database.

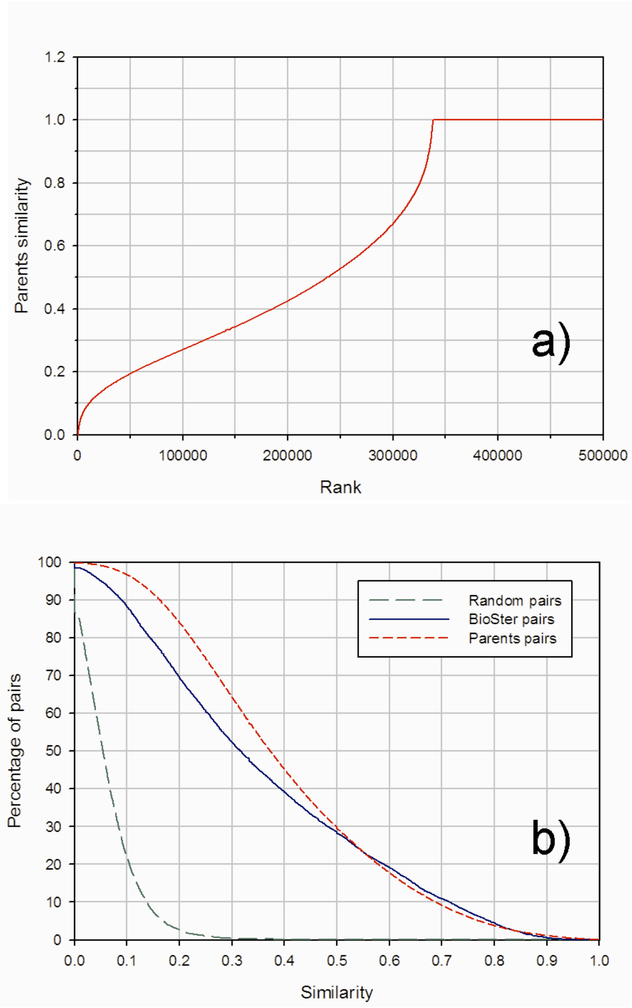


Figure 6. A) Comparison of the similarities of parents obtained from two independent iterative AP clustering runs starting from two random orderings of the same dataset. B) Distribution of LINGOsim values between parents assigned to the same molecule in two independent iterative AP clustering runs. The similarity between parents assigned to the same molecule in two runs is higher than the similarity distribution of bioisoster pairs in the BIOSTER database (http://accelrys.com/products/accord/chemical-databases/bioster.html). The similarity distribution of random pairs of molecules is shown as a comparison.

Eigenmolecules of the PubChem chemical space

After iterative AP clustering, each ancestor is hierarchically connected to a certain number of molecules in the database, which we call its branching number. Figure 7a shows the accumulated number of molecules associated with different ancestors ordered by their branching number. 47.2 % of the molecules in the database are associated with only 10 ancestors. In contrast, 52 of the 101 less branched ancestors (54.4%) collectively account for only 0.03% of the database.

Figure 7b shows the branching number of the ancestors, highlighting a clear distinction between the first more highly branched ancestors and the rest. The 49 ancestors with the highest branching. can be considered “eigenmolecules” of the database. They constitute a small set of highly dissimilar (“orthogonal”) molecules that can be linked by similarity to most of the database. Given the degeneracy of clustering arrangements discussed above, the molecules forming an eigenmolecule set may change depending on the clustering experiments. The structures of eigenmolecules obtained in a representative iterative AP-clustering experiment are provided as supplementary information.

The average cluster size generated by the AP clustering algorithm with the setting used is around 10 and therefore, four levels of clustering generate around 50 ancestors from 500,000 molecules, which is close to the value found. In contrast, scarcely branched ancestors and singletons correspond to rare structures. Their number is expected to be proportional to the size of the database and remains constant at each clustering level. We propose that four-level clustering is a suitable level to generate sets of ancestor s of the PubChem database formed by an approximately similar number of highly branched and unique structures.

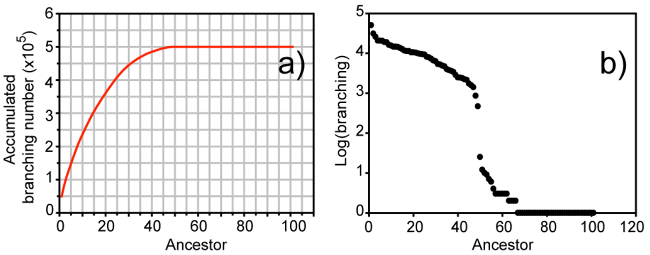


Figure 7 Ancestor branching. Ancestors (top level exemplars) are ordered according to their branching number. **a)** Accumulated number of database molecules hierarchically connected to different ancestors. **b)** Branching number of individual ancestors.

The set of molecules that are connected to eigenmolecules represent a filtered database, in which rare structures have been removed. A set of PubChem entries identified as low branching ancestors is given as supplementary material. These infrequent structures include hypothetical structures or representation errors. Iterative AP clustering of a filtered database in which the low branching ancestors and their related molecules had been removed gives 54 eigenmolecules with branching larger than 10^3 (see supplementary material).

Raghavendra and Maggiora [20] proposed the use of small basis sets of molecules and a generalized Fourier analysis method to represent larger databases. Molecules are assimilated to vectors in an abstract multidimensional space. Although the explicit coordinates of a molecular vector are not defined, the inner product is assimilated to the similarity between the molecules. By using a small number of molecules in the basis set, a representation error is introduced. This error can be quantified, as described in the methods section. If all the molecules of the database were included in the basis set, the error would be zero. As the number of molecules in the basis set decreases, the error increases. Thus a proper selection of the molecules included in the basis sets will decrease the error of the representation. An optimized set would include the smallest number of molecules that capture the structural characteristics of most of the database.

We compared basis sets of different sizes formed by including molecules from either: i) the set of 49 eigenmolecules; ii) a set of randomly chosen molecules; or iii) a set of random molecules with the same molecular weight distribution as the eigenmolecules. The maximum basis set size is restricted to 49 molecules. The incompleteness error was computed using a test set of 4,000 molecules chosen randomly from PubChem (excluding molecules in the 500,000 training set from which the eigenmolecules were defined).

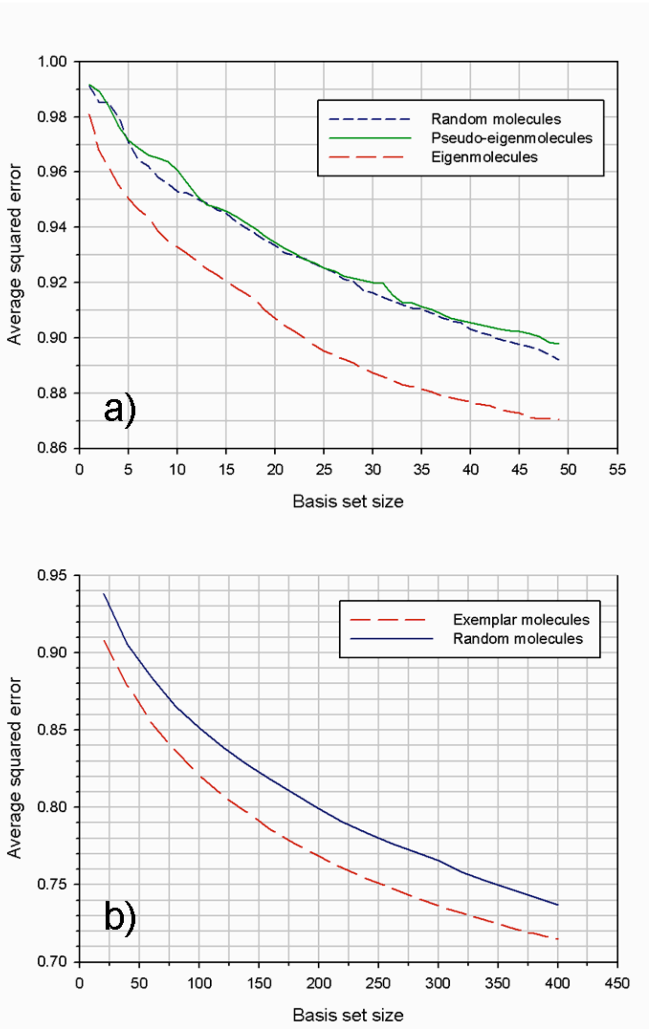


Figure 8. Incompleteness error of molecular basis sets generated from cluster exemplars or randomly chosen molecules. a) Eigenmolecules (the 49 ancestors with larger branching numbers, ordered by their branching number). b) Level 3 clustering exemplars (great-grandparents).

Figure 8a compares the eigenmolecule set with the two different random sets. The error decreases as the basis set size increases. While the number of eigenmolecules is restricted to 49 in this case, larger basis sets can be obtained using level 3 exemplars instead of eigenmolecules.

Figure 8b compares basis sets formed by level 3 exemplars or by random molecules. Depending on the error level, the required number of exemplars is 30-50% smaller than the number of random molecules.

The incompleteness error provides a quantification of the capacity of a small dataset to represent a larger database. Using this metrics, eigenmolecule sets are shown to be a good molecular representation of the PubChem dataset.

The most important characteristics of eigenmolecules are topological: eigenmolecule sets are highly diverse and their members are connected to a large proportion of the database using a small number of connection steps.

To evaluate the connectivity of the eigenmolecule set in the context of a larger subset of PubChem, 16 million molecules from this database were scanned to select those with similarity over a threshold value to the 49 eigenmolecules. The threshold was selected for each ancestor so that only the top 5% was retained when all molecules linked to a given ancestor in the 500,000-molecule set were ranked by their similarity to this ancestor. To compute the similarity, only the LINGOs present in the ancestors were considered. This approach provides a set of 2,269,570 molecules, which contains no rare molecules or singletons. A similar selection carried out using a random set of 49 molecules and a threshold equal to the average value used in the ancestor-based selection, selects only 672,541 molecules. This observation indicates that the eigenmolecule set is connected to a larger number of molecules not only in the “training set” of 500,000 molecules used to derive it but also in the complete PubChem database.

Application of iterative AP clustering to a set of 500,000 molecules selected from the highly connected 2,269,570 set gave only 39 highly branched ancestors. This result shows that the number of eigenmolecules is related to the diversity of the database.

Clusters group molecules with similar physical properties and biological activity.

Figure 9 shows the correlation between the intra-cluster average value of distinct physical properties and the corresponding values of the cluster exemplar. Properties were calculated from IUPAC names as previously described..[16]

The excellent correlations observed ($R^2 = 0.90$) demonstrate that the clustering of the database based on LINGO similarities captures relevant target-independent physical and pharmacological properties. This finding is consistent with previous demonstrations of the capacity of LINGO-based methods to provide accurate estimates of global properties. [14, 16]

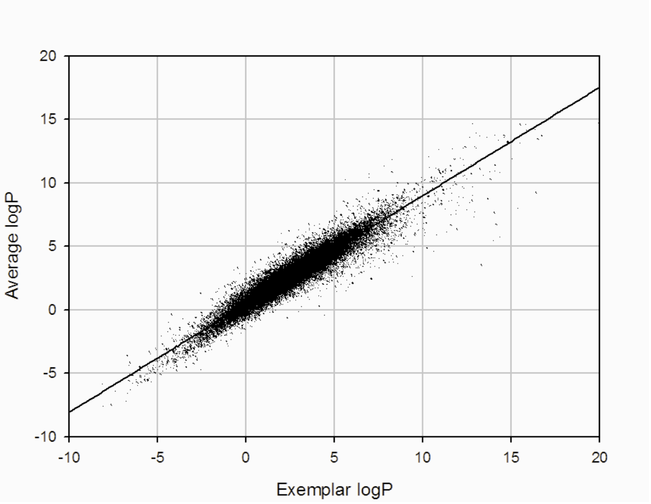


Figure 9 Correlation between logP, of exemplars and the corresponding intra-cluster averages using a 460.000 molecule subset of PubChem. Clustering was based on the SMILES representations. logP was calculated from IUPAC names[16]

In order to demonstrate that LINGO-based clustering also captures biological properties, such as ligand binding to specific targets, we used the DUD database, which contains validated ligand sets for 40 targets. [22] The list of targets present in the DUD database is given as supplementary information. Each ligand is associated with a number of non-ligands or “decoys” that share some physical properties with the ligand. Discrimination between true ligands and decoys provides a strict benchmark for virtual screening that discriminates genuine topological-structural features of the ligand-target interaction from mere physical constraints on the ligand properties.

We performed LINGOsim-driven iterative AP clustering of the complete merged DUD database (ligands and decoys for all targets). The distribution of ligands of a particular target in distinct clusters was compared with the distribution of the same number of decoys (randomly selected from the set associated with the same target) and the same number of molecules randomly chosen from the complete DUD database.

The Cluster Specificity Index (CSI, see methods section) provides a measure of the enrichment of ligands in particular clusters. When the total number of clusters is very high, randomly chosen molecules tend to fall into different clusters, giving a CSI close to one. Larger CSI values indicate an enrichment in particular clusters.

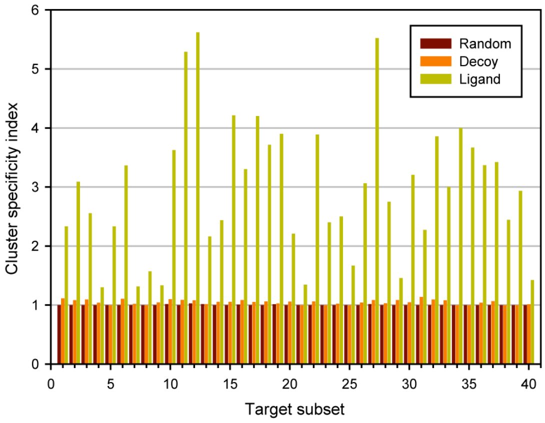


Figure 10. Enrichment factors for ligand sets of distinct DUD targets at the first clustering level. The number of clusters occupied by equivalent sets of ligands, decoys and random molecules are compared. The sets corresponding to different targets are normalized. The plots corresponding to clustering levels are shown as supplementary information.

The CSI values for the 40 ligand sets applied to the lowest clustering level are shown in Figure 10. Ligands of the same target are enriched in a smaller number of clusters than equivalent numbers of molecules that do not share a common biological activity (random set) even if they have similar physical properties (decoy set). The decoy sets show CSI values slightly higher than the random sets, thereby indicating that physical properties make a small contribution to the clustering of ligands from the same target. However, the large difference between true active ligands and decoys shows that the structural characteristics responsible for the specific biological activity of ligand sets have been captured by LINGOsim and correctly clustered by the iterative clustering algorithm.

Figure 11 shows the distribution of COX2 and EGFR ligands in the DUD database ordered by iterative AP-clustering. The ligand frequency is the number of active ligands per bin divided by the bin size. The 348 COX2 ligands appear in three peaks that group 186, 73 and 14 ligands. The EGFR ligands are more diverse but the bins contributing to the main peak are formed almost exclusively by EGFR ligands. The separation of ligand classes according to their LINGO similarities is reminiscent of a chromatographic process in which separation based on physical properties allows the isolation of molecules with a given biological activity and we suggest to call the LINGO based ordering of a database as LINGO-based virtual chromatography.

4 Conclusions

The usefulness of currently available large databases depends on the capacity to structure the represented chemical space and navigate through it. LINGO-based similarity provides a highly efficient method to generate the required large similarity matrices that capture the structure of chemical space. We corroborate that LINGO methods discriminate between bioisosteric and random pairs and that they can also be used to predict similarities between physical properties.

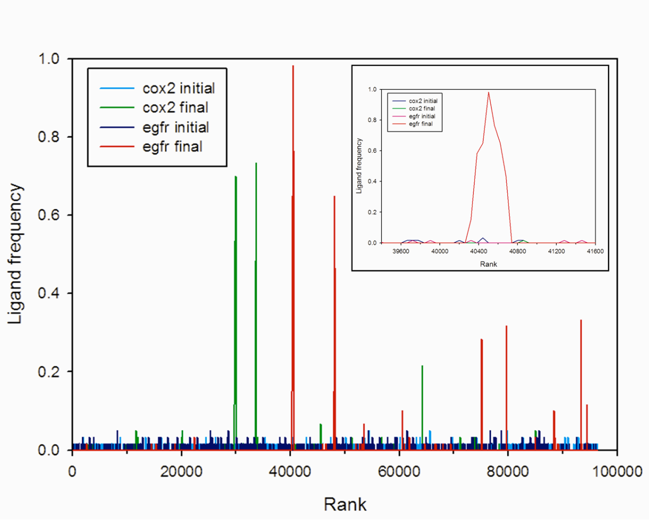


Figure 11. Virtual LINGO chromatography showing the distribution of COX2 and EGFR ligands in the ordered DUD database. The inset shows an expansion of the main EGFR peak. Bin size is 60.

AP clustering provides a robust and very effective clustering algorithm well suited to the large databases and non-metric distances found in chemical applications. Compared to other clustering algorithms, AP clustering is intensively data-driven, the predefinition of the number of clusters is not required and the cluster centers are chosen to be optimal exemplars (stay in the center) of the defined clusters, thus allowing iterations to generate a hierarchical clustering structure. The computational cost increases linearly with the number of similarity values and is therefore very well suited for the clustering of very large datasets. The algorithm easily converges for datasets as large as 3,000 molecules but cannot be applied directly to 500,000 molecules. We have developed an iterative procedure that generates a stable hierarchical clustering structure for the complete set starting from smaller random subsets. This structure is associated with a reordering of the database so that similar molecules are placed next to each other. The procedure converged after 25 iterations, as shown by the average similarity between groups of adjacent molecules and by comparing the composition of the clusters. Degenerate clustering arrangements can be obtained but it was demonstrated that the exemplars are either identical or have a similarity distribution equivalent to that found in bioisosters. Thus a functionally stable clustering arrangement is obtained.

After four levels of clustering, the highest level exemplars included two sharply divided sets of approximately the same size. The first set included what we called “eigenmolecules”, which collectively were linked by similarity to more than 99% of the database. The second set was formed by singletons or rare structures. Sets of exemplars generated by hierarchical AP clustering were shown to provide a much better representation of the complete dataset than the same number of random molecules using the generalized Fourier method described by Raghavendra and Maggiora.

The combination of LINGOsim and the AP algorithm generates clusters that include molecules with similar physical properties. Using the DUD database, we found that molecules known to bind to the same biological targets were enriched in particular clusters.

The literature includes a large number of clustering methods as well as similarity measures. AP clustering and LINGOsim are among the most effective methods and are well suited for large datasets. Iterative AP clustering based on LINGOsim has the capacity to generate a target-independent structure of the PubChem database. This new approach clusters ligands for the same targets and, therefore, could be used to speed up similarity-based virtual screening. Work is in progress to compare the performance of the proposed AP clustering-LINGOsim method with other widely used clustering methods and similarity measures.

Acknowledgements

We thank David Vidal (Chemotargets SL) for his contribution to the early development of LINGO tools and Prof. Peter Willett for reading a preliminary version of the manuscript and useful comments. This work was partially supported by funds from the Spanish "Ministerio de Educación y Ciencia" (BIO2007-63458) and the "Generalitat de Catalunya" (2009 SGR 1352). GC holds a predoctoral fellowship from the "Ministerio de Educación y Ciencia".

References

[1] O.A. Raevsky, S.V. Trepalin, H.P. Trepalina, V.A. Gerasimenko, O.E. Raevskaja, *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 540–549.
[2] H. Waterbeemd, E. Gifford, *Nature Rev. Drug Discov.* **2003**, *2*, 192-204.
[3] M.A. Johnson, G.M. Maggiora, in *Concepts and Applications of Molecular Similarity.* (Eds.: John Wiley and sons) New York, **1990**.
[4] S.J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, P. Baldi, *Bioinformatics* **2005**, *21*, i359-i368.
[5] P. Willett, J.M Barnard, G.M Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
[6] F.L. Stahura, J. Bajorath *Curr. Pharm. Des.* **2005**, *11*, 1189-1202.
[7] Y.C. Martin, J.L. Kofron, L.M Traphagen, *J. Med. Chem.* **2002**, *45*, 4350-4358.
[8] G. M. Downs, J. M. Barnard, *Reviews in Computational Chemistry*, **2002**, 18, 1-40, Wiley-VCH, John Wiley and Sons, Inc
[9] B.J. Frey, D. Dueck, *Science* **2007,** *315*, 972-976.
[10] G.M. Maggiora, V. Shanmugasundaram, J. Bajorath, in *Molecular Similarity Measures. Cheminformatics: Concepts, Methods and Tools for Drug Discovery.* (Eds. Humana Press) Totowa, New Jersey, **2004**, pp 1-50.
[11] A.R. Leach, V.J. Gillet, in *An Introduction to chemoinformatics*, (Kluwer Academic Publishers), Dordrecht. **2003**.
[12] A. Bender, R.C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.
[13] N. Nikolova, J. Jaworska, *QSAR Comb. Sci.* **2004**, *22*, 1006-1026.
[14] D. Vidal, M. Thormann, M. Pons, *J. Chem. Inf. Model.* **2005**, *45*, 386-393.
[15] D. Vidal, M. Thormann, M. Pons, *J. Chem. Inf. Model* **2006**, *46*, 836-843.
[16] M. Thormann, D. Vidal, M. Almstetter, M. Pons, *The Open Applied Informatics Journal* 2007, *1*, 28-32.
[17] G.A. Grant, J.A. Haigh, B.T. Pickup, A. Nicholls, R.A. Sayle *J. Chem. Inf Model.* **2006**, *46*, 1912-1918.
[18] T.I. Oprea, J. Gottfires, *J. Comb. Chem.* **2001**, *3*, 157-166.
[19] C. Klein, D. Kaiser, S. Kopp, P. Chiba, G.F. Ecker, *J. Comp. Aid. Mol. Des.* **2002**, *16*, 785–793.
[20] A.S. Raghavendra, G.M. Maggiora, *J. Chem. Inf. Model.* **2007**, *47*, 1328-1340.
[21] http://www.chemaxon.com/
[22] N. Huang, B. Shoichet, J.J. Irwin, *J. Med. Chem.* **2006**, *49*, 6789-6801.

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff)