Molecular diversity and its analysis

Dominique Gorse, Anthony Rees, Michel Kaczorek and Roger Lahana

We have recently developed a novel strategy for the rational design of compounds. This '*in silico* screening' approach is based on the design and screening of virtual combinatorial libraries. Screening is performed using defined rules derived from a comprehensive description of active and inactive molecules in a relevant learning set. This strategy allows the development of potential ligands without the necessity of any knowledge of the 3D-structure of the target receptor. Key to the success of such methods is the quality of the information being processed, in particular, the diversity of the data in the context of the molecular population in the libraries concerned. Here, we review the problem of data diversity, its definition and its analysis using a new software tool, named Diverser.

Biological activity is usually the result of the interplay of a number of complex processes, which cannot be easily represented by a set of linear relationships. To describe these processes better, non-linear variable mapping can be used, where the activity is represented by a non-linear function of structural, topological and molecular descriptors[1]. Analysis of such descriptors for active versus inactive compound sets can allow the definition of filters or 'activity codons' that differentiate between the two classes of molecules. This contrasts somewhat with the more common medicinal chemistry, or structure-based design approaches in that here, the inactive molecules are as important to the process of defining a filter as the active molecules.

A filter is defined by the range of variation of a given descriptor for all known active compounds when compared with the range of variation of the same descriptor for all known inactive compounds. This is intrinsically a non-linear process because the majority, if not all, of the descriptors are independent of one another. By contrast, structure-based design techniques can only be effectively applied when the receptor is known, its structure is known, the potential ligands are preferably not too flexible and the ligands are structurally similar. Because the majority of pharmaceutical targets do not often conform to all these requirements, alternative strategies are required. Flexibility is a *sine qua non* of biological activity. Hence, any rigorous method of drug discovery will incorporate some measure of flexibility, either into the target or the ligand, or both. Thus, instead of the optimization process operating on a prescribed conformation of a molecule, or on a given set of conformations, the process will operate on the conformational landscape of the molecule representing its entire conformational phenotype[2].

Standard combinatorial chemistry can provide hits. However, these hits are of little use if they are structurally unrelated because leads cannot be easily derived from them through 'series-driven' medicinal chemistry. The probability of finding related hits is, among other factors, related to the diversity of the starting library. The question is, therefore, how to go about analysing the diversity of the virtual combinatorial library generated at the beginning of the process. Typical virtual libraries can contain millions, if not billions, of compounds. However, it is likely that many members of such libraries will not significantly

Dominique Gorse*, **Anthony Rees**, **Michel Kaczorek**, **Roger Lahana**, Synt:em, 145 Allée Charles Babbage, 30000 Nîmes, France. Anthony Rees is also in the Dept of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK BA2 7AY. *tel: +33 466 048 666, fax: +33 466 048 667, e-mail: dgorse@syntem.eerie.fr

1359-6446/99/$ – see front matter © Elsevier Science. All rights reserved. PII: S1359-6446(99)01334-3

differ from one another in terms of their descriptor profiles. The quantitative scoring of the diversity of every molecule in the library will enable the screening out of redundant information and the definition of sub-libraries having the same diversity as the initial library, but with a significantly lower number of members. At the same time, diversity voids will be detected that, when appropriate, can be filled in by new compounds.

Choice of the molecular descriptor space

2D descriptors have been shown to be very effective in discriminating classes of molecules[3]; in addition, they are fast to compute and they are, by nature, conformation-independent. Thus, their application to large collections, or libraries, of molecules is efficient and economical with computer time. 2D descriptors are also independent of the flexibility of the molecule. In the diversity approach described in this article, virtual molecules are represented by such a set of conformation-independent descriptors. As indicated earlier, molecules can also be quantitatively analysed in terms of their conformational landscapes, but this type of analysis relates more to the process of drug discovery and optimization, and is outside the scope of this review.

Descriptors define what can be thought of as a 'chemical space' that has as many dimensions as descriptors. These descriptors comprise a variety of topological descriptors, such as, for instance, the standard connectivity and shape indices[4–6], and conformation-independent molecular properties, such as group counts, molecular weight, and so on. An example of a typical set of descriptors used in a real analysis can be found in Ref. 7, which shows how a rational process based on such descriptors has led to the discovery of new immunosuppressive molecules exhibiting an *in vivo* activity more than 100 times higher than the initial lead compound.

Hundreds of descriptors can be generated for a given learning set of molecules, but only a few of them are kept to describe the diversity of a whole database. The reason for this is that, ideally, descriptors should not be intercorrelated. Rather, they should have a normal or uniform distribution, they should be relevant to the particular problem under study and they should be reasonably fast to calculate.

The first step in selecting descriptors is to remove those that are significantly intercorrelated ($r > 0.6$), those that exhibit no variance over the data set (single valued) and those that are highly discrete. Applying these rules of common sense leads typically to the withdrawal of some 80% or more of the initial set of descriptors. The distribution of molecules according to each descriptor is then analysed using median and mean comparisons, skewness and kurtosis coefficients, variation coefficient and χ^2 test. Standard transformations by square root or log functions can be applied to reach normal distributions. Finally, a choice between descriptors of equivalent statistical characteristics is based on their chemical interpretability and their computational complexity. The resulting space of molecular descriptors would then often consist of fewer than 20 dimensions.

Diversity analysis

Virtual libraries of molecules often contain a very high number of compounds (typically between 10^5 and 10^9), which are represented by a significant number of descriptors (typically a few dozen). This places a restriction of the type of data analysis that can be applied to such datasets. For example, popular clustering techniques, such as the Ward or Jarvis–Patrick techniques, are of order $O(N^2)$, which is acceptable for up to ~10,000 data points, but certainly not for higher orders of magnitude, where $O(N)$ becomes critical. To address the problem of processing large virtual libraries, we have developed new algorithms that focus on the computing time problem, the memory requirements of the process, the various ways of inputting variables and the analysis of results.

Encoding the descriptors

In Diverser, a descriptor is coded as an n-bit key and a molecule being described by a set of descriptors is represented by what can be called a 'molecular fingerprint', which is the concatenation of as many n-bit keys as descriptors for a given molecule. A database fingerprint is defined as the summation (logical OR) of a set of molecular fingerprints. Descriptor values are encoded in the n-bit key by a process called data coded by location, or DCBL (Fig. 1), which has some similarities to the diverse property-derived (DPD) code[8]. Two situations can be defined depending on the nature of the descriptor. (1) For continuous values of the descriptor (Fig. 1a), each bit represents a range of values. The increment between bits is determined by the difference between the maximal and minimal descriptor values, divided by the number of bits of the key. The number of bits is defined by the user. Increasing the number of bits results in an increase in the resolution of the descriptor. (2) For discrete values of the descriptor (Fig. 1b), each bit represents a set of specific values. Thus, the number of bits is determined by the discrete increment between bits, which is in turn defined by the user.

When loading a database into Diverser, DCBL codes are generated and the database fingerprint is set up (Fig. 2).

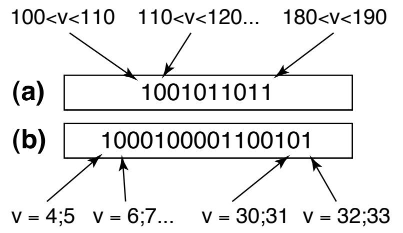


Figure 1. *Data coded by location (DCBL) coding scheme (a) for continuous values, (b) for discrete values.*

Approximately six minutes are needed to load and to encode 10^6 molecules described by 13 descriptors (with a Silicon Graphics R10000 processor). All DCBL codes are held in the memory as integer values for faster access. Moreover, each bit of the database fingerprint points to the number of molecules that belong to the corresponding range of values. This cardinal information allows quick and easy examination of the distributions of the properties (Fig. 2) and, as discussed later, can be used to reduce computational time during the diversity analysis.

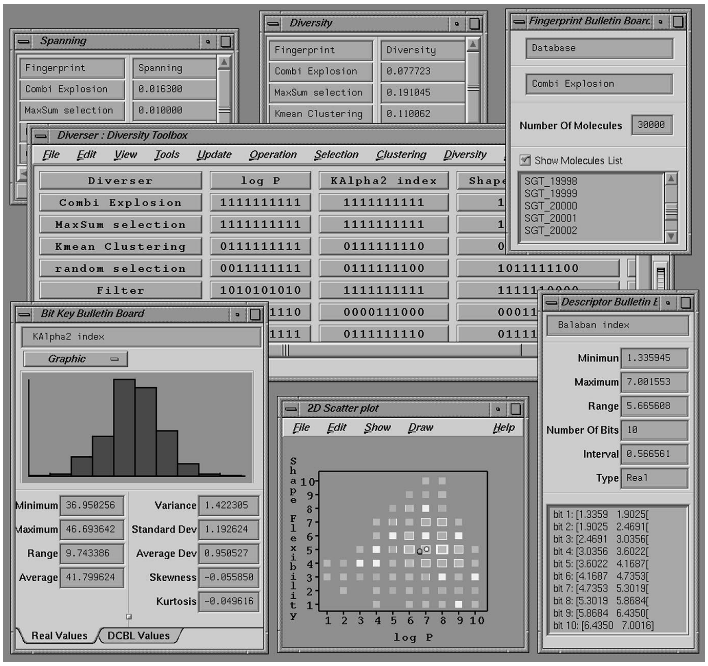
Similarity-based and grid-based methodologies

In a recent review, Van Drie and Lajiness[9] point out the differences between similarity-based and grid-based methods for the selection of molecules and evaluation of their diversity. Diverser has been designed to combine the two approaches. Partitioning methods consist of dividing the descriptor space into cells and then attributing molecules to their respective cells. The choice of the number of cells used in the analysis is a sampling issue[10] – the number of cells should certainly be less than the number of molecules. Cummins *et al.*[10] have addressed this issue and suggest the use of two molecules per cell. According to their rule of thumb, the optimal number of bits per descriptor is $b = (N/2)^{1/d}$, where N is the number of molecules and d is the number of descriptors. In the present example, databases are of the order of a million molecules, described by approximately ten highly non-correlated descriptors, hence the number of bits per descriptors is usually less than five. The main benefits of partitioning methods[8–10] are the following:

- fast and easy database comparisons
- possibility of estimating database diversity
- straightforward identification of diversity voids
- direct selection of a subset

As indicated above, an increase in the number of bits leads to a better resolution of the descriptor. For example, if the number of bits of a descriptor is ten, then the resolution between two bits, or classes, is 10% of the difference between the maximal and minimal values of that descriptor. If the number of bits is 20, then the resolution is 5%, and so on.

The DCBL coding can be seen as a data standardization of classes. The DCBL codes of a molecule can represent either the cartesian coordinates or the vector coordinates of the molecule in the descriptor space. From a



Fingerprint	Spanning
Combi Explosion	0.016300
MaxSum selection	0.010000

Fingerprint	Diversity
Combi Explosion	0.077723
MaxSum selection	0.191045
Kmean Clustering	0.110062

Diverser	log P	KAlpha2 index	Shape
Combi Explosion	1111111111	1111111111	1
MaxSum selection	1111111111	1111111111	1
Kmean Clustering	0111111111	0111111110	0
random selection	0011111111	0111111100	1011111100
Filter	1010101010	1111111111	111111
	1110	0000111000	00011
	1111	0111111110	01111

Database

Combi Explosion

Number Of Molecules 30000

☑ Show Molecules List

SGT_19998
SGT_19999
SGT_20000
SGT_20001
SGT_20002

KAlpha2 index

Minimum	36.950256	Variance	1.422305
Maximum	46.693642	Standard Dev	1.192624
Range	9.743386	Average Dev	0.950527
Average	41.799624	Skewness	-0.055850
		Kurtosis	-0.049616

Balaban index

Minimun	1.335945
Maximum	7.001553
Range	5.665608
Number Of Bits	10
Interval	0.566561
Type	Real

bit 1: [1.3359 1.9025[
bit 2: [1.9025 2.4691[
bit 3: [2.4691 3.0356[
bit 4: [3.0356 3.6022[
bit 5: [3.6022 4.1687[
bit 6: [4.1687 4.7353[
bit 7: [4.7353 5.3019[
bit 8: [5.3019 5.8684[
bit 9: [5.8684 6.4350[
bit 10: [6.4350 7.0016]

Figure 2. *Main windows used in the Diverser program. The middle window displays fingerprints from databases, filters, selection results or comparisons. Bulletin boards allow monitoring of the main characteristics of: a selected bitkey, which is the intersection of one descriptor and one fingerprint (bottom left window); a selected descriptor (bottom right window); and a selected fingerprint (top right window). Other windows are used to display distibutions or results from diversity evaluation.*

computational point of view, the number of dimensions of the descriptor space has no limitation and, because distances between molecules in this space can be calculated, similarity- or dissimilarity-based analyses can be carried out.

Distances, dissimilarity and the DCBL code

Numerous distances and coefficients have been proposed and used for dissimilarity evaluation[11]. Among them, the Tanimoto coefficient has demonstrated its value in measuring intermolecular similarity[12–14]. However, this coefficient is more relevant for binary fingerprints than for molecular vectors. Intermolecular pairwise dissimilarity, or similarity, available in Diverser are based on the Euclidean distance, the squared Euclidean distance, the City-Block or Manhattan distance and the Cosine coefficient (Fig. 3).

The Euclidean distance between two molecules is the geometric distance in the multidimensional space, where x and y are two molecules, k is the dimension and x_k and y_k are the coordinates of molecule x or y for dimension k:

$$d(x,y) = \sqrt{\sum_k (x_k - y_k)^2}$$

The squared Euclidean distance is merely the sum for each dimension of the squared differences of the molecular coordinates:

$$d(x,y) = \sum_k (x_k - y_k)^2$$

This distance has the effect of giving more weight to molecules that are further apart. The City-Block or Manhattan distance is the sum of the coordinate differences over all the dimensions:

$$d(x,y) = \sum_k |x_k - y_k|$$

In general, this distance leads to results similar to those obtained with the Euclidean distance. With the Cosine coefficient, the similarity between two molecules is estimated by the cosine of the angle formed by the two corresponding molecular vectors. Dissimilarity is simply the reciprocal of the similarity. The choice of metrics obviously has an impact on the final results but, more importantly, it affects the computational time needed for diversity analysis. This can be increased dramatically when sums of pairwise dissimilarities are involved.

The use of the Euclidean distance does not allow any gain in algorithm speed. Calculation of the sum of all the pairwise dissimilarities has a time requirement of order $O(N^2)$. The calculation time of the sum of all the pairwise cosine coefficients also has an order of $O(N^2)$ but, as suggested by Peter Willet and co-workers[15,16], the summation is equivalent to the calculation of the dot product of a 'centroid' vector with itself, for which a computational time of order O(N) is expected. It should be pointed out that the choice of the origin of the molecular vectors used to construct the centroid is crucial, to avoid inducing a privileged orientation in the chemical descriptor space.

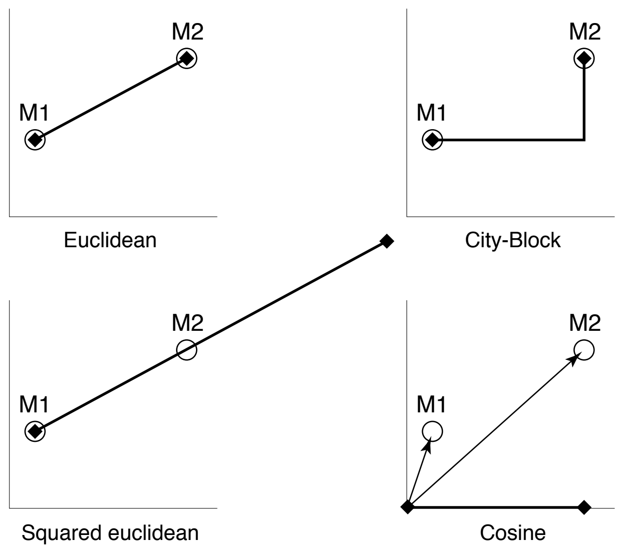


Figure 3. *Distances and association coefficient used in the Diverser program.*

Using the cardinal information encoded by the Diverser fingerprint also allows a rapid way of evaluating the sum of all the pairwise dissimilarities of one molecule with a set of N other molecules when squared Euclidean or City-Block distances are used. For these distances, molecular coordinate differences (or their squares) across dimensions that are common between molecules, can be factorized. The sum of all the intermolecular dissimilarities does not depend on the number of molecules but only on the total number of bits. For example, the computational time spent on a Silicon Graphics R10000 processor in calculating the sum of all the pairwise molecular dissimilarities of a million molecules described by 13 descriptors of 10 bits each are given below, assuming the database has already been loaded (six minutes):

- 55 days when using the Euclidean distance (this is an extrapolated value)

- almost instantaneous when using the City-Block distance associated with the fingerprint cardinal information
- 20 seconds when using the Cosine coefficient with the centroid dot product.

Dissimilarity-based selection

Compound selection is one of the major issues in diversity analysis. Selection is important for two reasons: first, to get a subset that is the most diverse or the most representative for a given biological assay; second, to get a subset that will enrich the diversity of another existing database. Among similarity and dissimilarity-based selection methods, clustering and maximal dissimilarity methods are implemented in Diverser. Using these two types of method, different types of answer are actually obtained. With maximal dissimilarity, a representative subset of a large database is obtained by selecting molecules that are as dissimilar as possible. When using clustering techniques, the subset consists of representative compounds from clusters of similar molecules.

Maximal diversity selection

Given one of the metrics previously described to estimate pairwise intermolecular dissimilarity, the dissimilarity between one molecule and a set of N molecules can be evaluated[17], either by the sum of the pairwise intermolecular dissimilarities (Sum) or by taking the minimal value found amongst all the pairwise intermolecular dissimilarities (Min). A comparison of these different ways of measuring dissimilarity in compound selection methods has been reviewed recently[14]. The authors tested a panel of dissimilarity-based methods and suggest that the maximal dissimilarity-based method using the Min dissimilarity (referred to as the MaxMin selection method) is effective and efficient in selecting compounds associated with a range of biological activity classes.

A drawback with the selection methods based on maximal dissimilarity (MaxSum or MaxMin) is that the expected time complexity is of order O(n^2N) when selecting n molecules from N. However, for the MaxMin selection, Polinsky *et al.*[18] have suggested an algorithm of order O(nN). As with the diversity evaluation, the use of the Cosine coefficient (centroid algorithm)[15,19] or the City-Block distances or the squared Euclidean distances combined with the fingerprint cardinal information in MaxSum reduces the time requirement to O(nN). These techniques can be applied to large databases. For example, selecting a subset of 200 (respectively 2000) molecules from a million on a Silicon Graphics R10000 processor would take around 35 minutes (respectively six hours), using either City-Block distance or centroid dot product.

K-means selection

The K-means clustering algorithm is a method of choice for large databases because the calculation-time requirement is of the order of O(nN). Such an application has been recently proposed by BCI (Refs 20,21) for compound selection. K-means belongs to a large family of algorithms that exhibit many variants in the exact details by which partitions are generated and adjusted. A better generic name might be 'mobile centres techniques'. The main stages of the mobile centres clustering method[22] are as follows (see Fig. 4):

- Step 1: K initial centroids are selected. The number of centroids K corresponds to the number of desired clusters.
- Step 2: Clusters are constructed by assigning each molecule to the closest centroid.

Then an iterative process is started:

- Step 3: Centres of gravity of the clusters are calculated, leading to new centroids.
- Step 4: Clusters are reconstructed by assigning each molecule to the closest centroid and Step 3 is repeated.

Iterations are stopped either when two successive iterations lead to the same partition, when intra-classes variance converges, or when a given fixed number of iterations has been carried out.

The K-means algorithm proposed by McQueen[23] differs from the 'mobile centres' method in its more efficient use of information. The centroid update (assigning molecules to a cluster, then recomputing the centroid) is applied at each step in the initial partitioning and during the iterations. The mobile centres or centroids always correspond to the centres of gravity, or mean, hence the name 'K-means'. Because it constantly updates the clusters, the K-means algorithm usually leads to a final partition in one iteration and, thus, it is much faster than the general mobile centres algorithm. However, the partition will depend on the order of the molecules in the database. Heuristics[20,21] can be used to gain significant improvements in computational performance. For example, if a cluster does not change between two iterations, it is not necessary to recalculate the distances between the molecules of that cluster and the centroids of other clusters that have also not changed. As shown in Fig. 5, the time spent for each iteration rapidly decreases as the centroids become stabilized.

Sensible choice of the original centroids is important because it will avoid the algorithm becoming trapped in local minima and will require fewer iterations to reach the global minimum. One of the most effective ways is to select the most dissimilar compounds using a MaxSum or MaxMin algorithm. As suggested by Diday[24], the definition of centroids that consist of one or more compounds on the basis of structural or chemical intuition, or for that matter any other 'common-sense' reason, will also be effective. When the initial centroids are selected randomly, different partitions from different sets of random centroids can be obtained, and strong-density areas found among these different partitions can be identified.

The most straightforward method of determining strong-density areas is to cross the K classes found in the P partitions and to sort the resulting K^P partition-product by decreasing size. The first partition-products in this sorting are the largest and most stable clusters found over the P partitions. Centroids of the first K partition-products can be calculated and a new partition created to assign molecules to these centroids. The main restriction lies in the value of K^P, whose value restricts the use of this method in practice to a small number of clusters (~100) and a few partitions (~3), far fewer than the number of clusters one would ideally like to evaluate (10 to 100 times larger). For this reason, rather than crossing the K classes from the P partitions together, it is preferable to cross them with a reference partition. Some strong-density areas are then highlighted, although their number is generally less than the classes obtained in each partition.

An important improvement would be to combine the 'mobile centres' algorithm with a hierarchical classification technique[25], such as the Ward method (Fig. 6). Mobile centres are then used as a pre- and post-processing of the Ward classification and the final partition is more effective and objective. The many variations that can be introduced in the mobile centres algorithm and in the combinations with other classification techniques opens up many

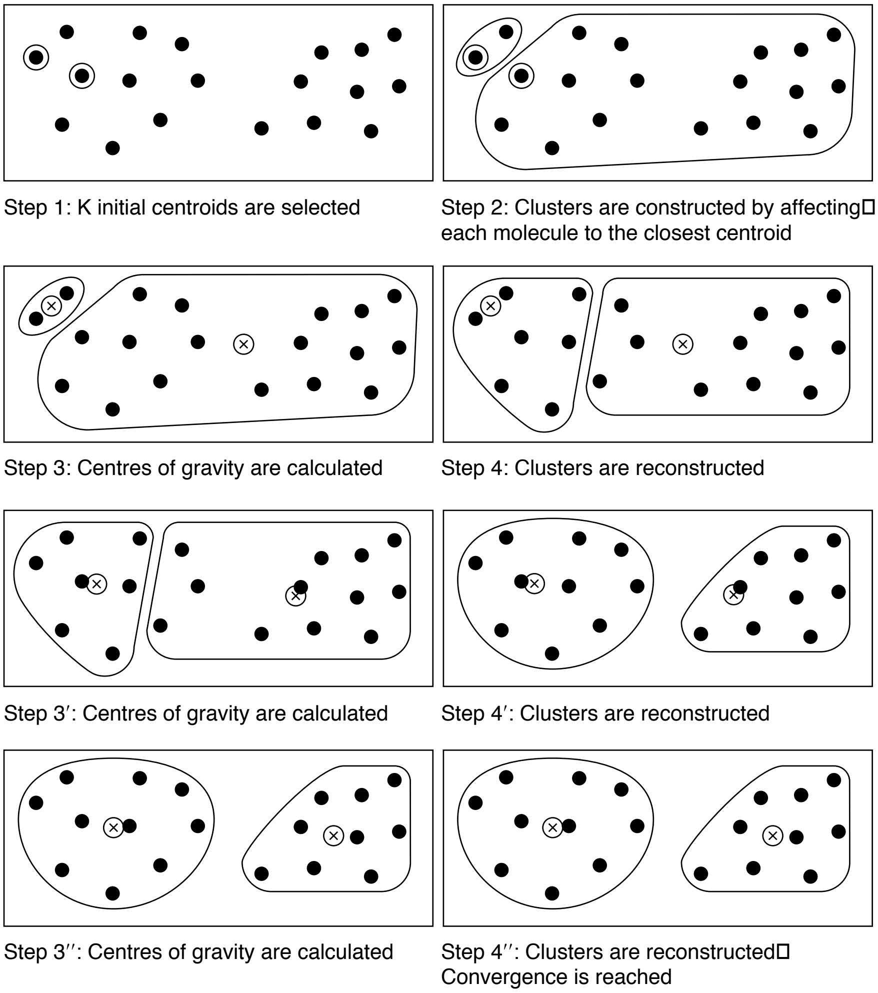

Step 1: K initial centroids are selected
Step 2: Clusters are constructed by affecting each molecule to the closest centroid
Step 3: Centres of gravity are calculated
Step 4: Clusters are reconstructed
Step 3′: Centres of gravity are calculated
Step 4′: Clusters are reconstructed
Step 3″: Centres of gravity are calculated
Step 4″: Clusters are reconstructed Convergence is reached

Figure 4. *The mobile centres algorithm.*

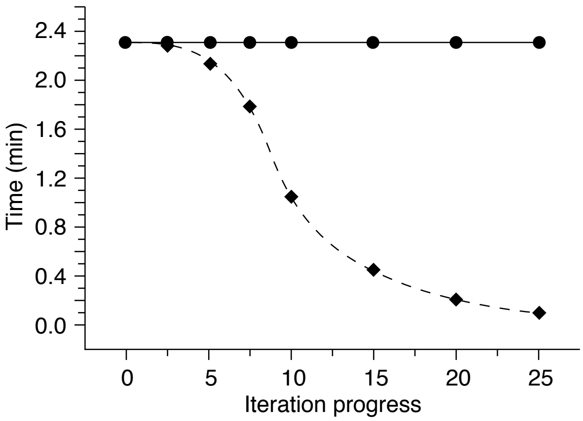


Figure 5. *The use of heuristics permits a decrease in the computational time as clusters tend to become stable.*

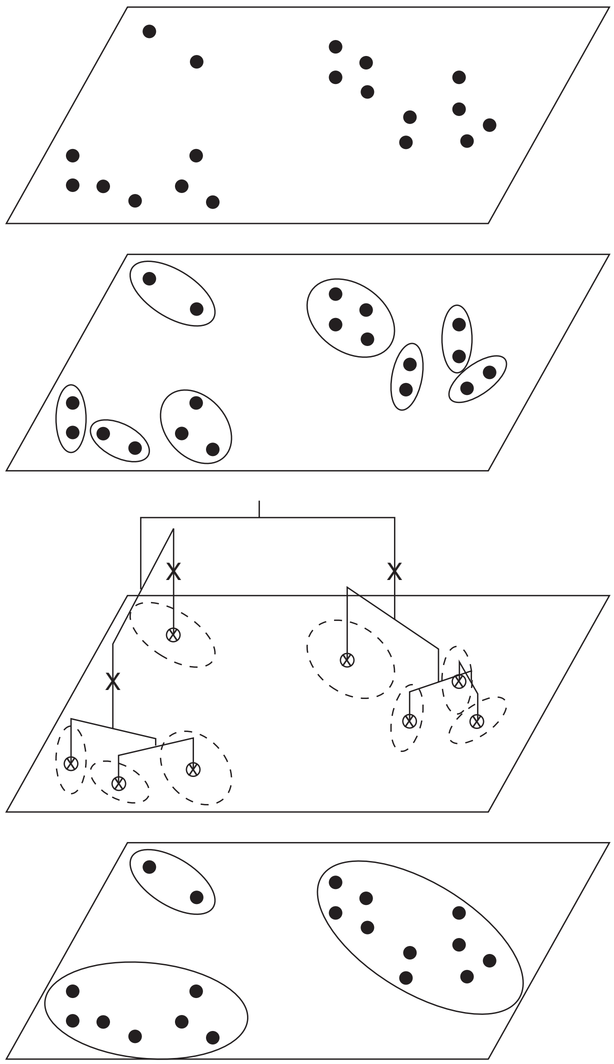


Figure 6. *Mobile centres used as a pre- and post-processing of the Ward classification.*

possibilities for the future development of efficient and effective clustering methods that can be applied to large combinatorial libraries.

Conclusion

The diversity of large compound databases and virtual libraries needs to be properly assessed. For example, it is important to be able to estimate and to compare overall diversity, to remove information redundancy, to select optimal subsets of the most diverse or the most representative molecules, and to identify unexplored regions of the chemical space. Diverser is a toolbox that can now satisfy all these requirements. Based on the DCBL code, it combines the benefits of fingerprint representation and space partitioning. Partitioning needs a reduced descriptor space. When more descriptors are used with large bitkeys, the DCBL code can be seen as a simplification of the raw data, allowing fast algorithms for maximal dissimilarity or mobile centres selections.

The choice of metrics and of the dissimilarity depends on the richness of the information desired, or needed, and on the computation time that can be applied. Using the cardinal information of the database fingerprints can significantly improve the calculation-time requirements for the selection of compounds and evaluation of their diversity. We expect these improvements to have an impact on the drug discovery process through, in particular, improved library design, leading to the discovery of entirely new molecular species as drug candidates.

Acknowledgements

The authors would like to thank David Turner and David Wilton from the University of Sheffield for helpful discussions on selection methods. We would also like to thank Geoff Downs from BCI Ltd for providing us details on their implementation of K-means technique.

REFERENCES

1 Grassy, G. *et al.* (1995) *J. Mol. Graph.* 13, 356–367

2 Grassy, G. and Lahana, R. (1993) in *Trends in QSAR and Molecular Modelling* (Wermuth, C., ed.), pp. 216–219, ESCOM Publishers

3 Brown, R. and Martin, Y. (1997) *J. Chem. Inf. Comput. Sci.* 37, 1–9

4 Ghose, A. and Crippen, G.M. (1986) *J. Comput. Chem.* 7, 567–577

5 Balaban, A.T. (1982) *Chem. Phys. Lett.* 89, 399–402

6 Hall, L.H. and Kier, L.B. (1992) *Reviews in Computational Chemistry* (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 367–422, VHC Publishers

7 Grassy, G. *et al.* (1998) *Nat. Biotech.* 16, 748–752

8 Lewis, R.A., Mason, J.S. and McLay, I.M. (1997) *J. Chem. Inf. Comput. Sci.* 37, 599–614

9 Van Drie, J.H. and Lajiness, M.S. (1998) *Drug Discovery Today* 3, 274–283

10 Cummins, D.J. *et al.* (1996) *J. Chem. Inf. Comput. Sci.* 36, 750–763

11 Willett, P., Barnard, J.M. and Downs, G.M. (1998) *J. Chem. Inf. Comput. Sci.* 38, 983–996

12 Barnard, J.M. and Downs, G.M. (1992) *J. Chem. Inf. Comput. Sci.* 32, 644–649

13 Brown, R.D. and Martin, Y.C. (1998) *SAR QSAR Environ. Res.* 8, 23–39

14 Snarey, M. *et al.* (1997) *J. Mol. Graph. Mod.* 15, 372–385

15 Holliday, J.D., Ranade, S.S. and Willett, P. (1995) *QSAR*, 14, 501–506

16 Turner, D.B., Tyrrell, S.M. and Willett, P. (1997) *J. Chem. Inf. Comput. Sci.* 37, 18–22

17 Holliday, J.D. and Willett, P. (1996) *J. Biomol. Screening*, 1, 145–151

18 Polinsky, A. *et al.* (1996) in *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery* (Chaiken, I.M. and Janda, K.D., eds), pp. 219–232, American Chemical Society

19 Pickett, S.D. *et al.* (1998) *J. Chem. Inf. Comput. Sci.* 38, 144–150

20 Downs, G.M. and Barnard, J.M., http://www.bci1.demon.co.uk

21 Downs, G.M. and Barnard, J.M. (1998) *Computational Approaches to the Design and Analysis of Combinatorial Libraries*, 14–16 April, University of Sheffield, UK

22 Forgy, E.W. (1965) *Biometrics* 21, 768

23 McQueen, J.B. (1967) in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press

24 Diday, E. (1971) *Revue Statistiques Appliquées* 19, 19–34

25 Wong, M.A. (1982) *J. Am. Stat. Assoc.* 77, 841–847

In short…

The gene therapy company, **IntroGene BV** (Leiden, The Netherlands) and the drug discovery company, **Tibotec NV** (Mechelen, Belgium) have embarked upon a joint venture and have formed a new company, Galapagos Genomics NV, which will be based in Mechelen, Belgium. Galapagos Genomics will focus on the identification of human gene function through the use of adenoviral technology developed by IntroGene, together with the high-throughput screening capabilities of Tibotec. The main role of the new company will be to create libraries of adenoviral vectors containing human genes to enable the study of protein products expressed in a real-cell environment.

Galapagos Genomics is to be led by Onno van de Stolpe as Managing Director, and the number of scientists employed by the company is expected to double by the end of the year. Mr van de Stolpe says that ‘We anticipate that our first validated library will be available for commercial screening by the end of this year, and subsequent libraries will follow thereafter’.

In the July issue of Drug Discovery Today…

Update – latest news and views

Convergent automated parallel synthesis
David G. Powers and David L. Coffen

Green fluorescent protein: applications in cell-based assays in drug discovery
Steven R. Kain

Dopaminergic agents for the treatment of cocaine abuse
Miles P. Smith, Alexander Hoepping, Kenneth M. Johnson, Monika Trzcinska and Alan P. Kozikowski

Intellectual property and chirality of drugs
I. Agranat and H. Caner

Monitor – new bioactive molecules, combinatorial chemistry, invited profile

Products