Intelligent *K*-means clustering in a multiple endpoint environment

E. Kolossov, IDBS; R.W. Stanforth, IDBS;
IDBS, 2 Occam Court, Surrey Research Park, Guildford, Surrey GU2 7QB, UK.
+44 1483 595000; ekolossov@idbs.com

Introduction

The main success of QSAR predictive technology is mainly related to the modelling of the single activity/property endpoint in chemical space. Unfortunately an optimisation of the targeted endpoint may result in losing other desired properties, for example, toxicity is increased or solubility reduced. Because of this, attempts have been made in the last few years to model chemical space against multiple endpoints. A good example of these attempts is the BioPrint Profile Prediction[1] based on General Neighbourhood Behaviour (GNB) modelling and Screener Sarileo[2], based on mapping between clusters in chemical space and biological fingerprints.

K-means clustering has proved successful in descriptor-based QSAR studies when used to cluster structures in descriptor space alone. We have also researched and implemented into the PredictionBase suite various ways of incorporating one activity variable into Intelligent *K*-means clustering. At its simplest, the activity may be included as if it were just another descriptor (albeit with a weighting to ensure that its influence is not lost if the number of descriptors is large).

Secondly, activity may instead make a regression-wise contribution, resulting in a hybrid of the straight *K*-means and regression-wise *K*-means algorithms. In this scheme each cluster is modelled by its centroid and its least-squares best-fitting model over the cluster, with distance-to-cluster being a combination of distance-to-centroid and prediction error according to the cluster's model.

A third scheme, which is a compromise between the computational speed of 'simple' activity contribution and the improved modelling power of full regression-wise *K*-means, uses simple incorporation of activity values, but preprocessed by expressing them relative to a fixed global regression line. It turns out that this scheme is equivalent to regression-wise *K*-means with a fixed model direction: only the constants are permitted to vary from cluster to cluster.

We now consider the situation upon moving into an environment with several activity variables (e.g. multiple endpoints). The first (simple contribution) and third (preprocessing relative to regression line) of the above activity contribution methods can carry over directly to the multiple endpoint case, (again with a suitable weighting to remove effects of varying relative numbers of descriptor/activity variables). Translating the regression-wise *K*-means contribution to a multiple endpoint environment is also possible, again by simply including separate regression-wise contributions for each activity. For a large number of endpoints, however, the computational complexity is compounded: for each iteration of *K*-means, the number of least-squares single-cluster regressions to be trained is the number of endpoints times the number of clusters.

Modelling Domain of Applicability

The intelligent *K*-means cluster algorithm[3] is a proven method for unsupervised classification of data in descriptor space. When applied to a dataset that has an underlying structure of several cleanly disconnected clusters in descriptor space, the algorithm can successfully find these clusters and classify the dataset accordingly.

More recently, we have been applying the intelligent *K*-means algorithm in a different way: we view the *K*-means clusters as modelling the shape of a dataset, rather than classifying it[4]. In this formulation, the clusters individually have little significance of their own, and their boundaries are somewhat arbitrary. However, taken as a collection, the clusters tile the region of descriptor space – the 'domain' – occupied by the dataset. The clusters therefore model the shape of the domain, regardless of whether the shape is simple or even convex.

Moreover, since the *K*-means clusters are roughly spherical, this cluster model of the shape of the domain can be expressed simply – and represented efficiently – as a collection of the cluster centroids and their radii. This is in stark contrast to *K*-NN (nearest neighbour) models of the domain, in which the representation consists of the entire dataset. In this sense a *K*-NN model is overfitted: it models the *noise* – the precise location of every data point – rather than the *trends* in the data scatter in descriptor space.

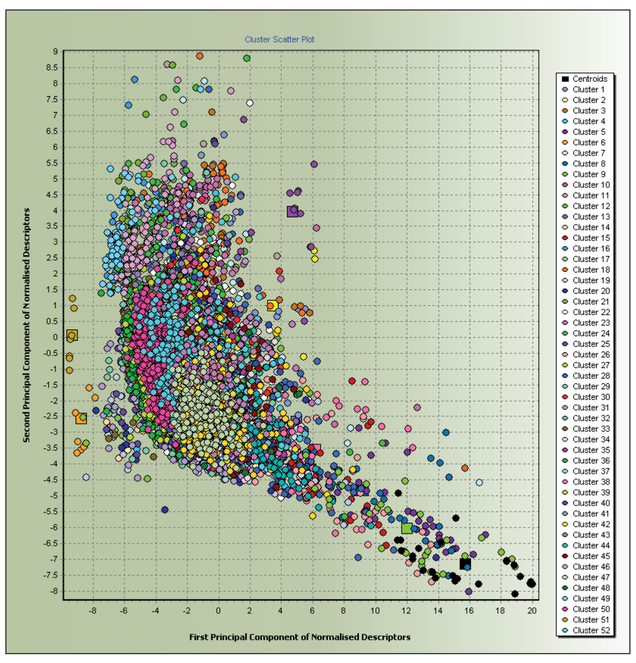


Figure 1. PredictionBase: *K*-means scatter plot with centroids

Cluster-Based Compositing: Distance to Domain

If a domain (of a dataset, or of applicability of a QSAR model) is modelled as a sphere in descriptor space, then the distance (of a test point) to that domain is measured as the Euclidean distance between the point and the sphere's centre. In order to allow meaningful comparisons, this distance can be expressed as a multiple of the sphere's radius; a distance greater than unity then corresponds to being outside the domain.

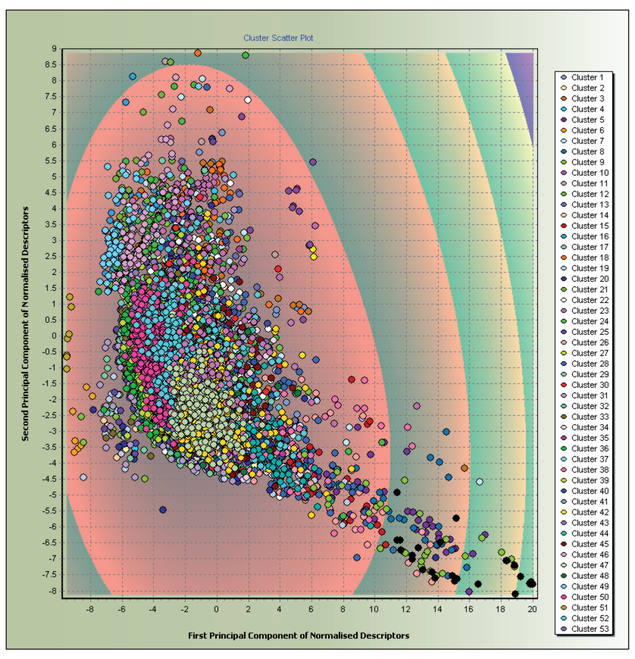


Figure 2. PredictionBase: *K*-means scatter plot with distance-to-domain contours

This distance notion can be extended in a straightforward fashion to domains modelled by *K*-means clusters as described above. A test point has a distance to each cluster, modelled by a sphere. Whereas the domain is considered as the aggregate composition of these clusters, so the distance (of the test point) to the domain should be taken as a weighted average of these distances to clusters.

In compositing distances to individual clusters into an overall distance to domain, the weightings in the average are chosen so that nearby clusters contribute the most and distant clusters have less influence. This can be viewed as a form of 'fuzzy clustering': the test point, rather than being treated as belonging to one cluster exclusively, is treated as having shared membership of all clusters to varying degrees.

In the particular fuzzy weighting scheme that we use, the fuzzy weighting $z_k(x)$ for cluster k (considering a test point x) is as follows[5]:

$$z_k(x) = (1/d_k(x)) \; [(1/d_1(x)) + \ldots + (1/d_K(x))]^{-1} \qquad (1)$$

where $d_k(x)$ is the distance (of test point x) to cluster k. The resulting distance-to-domain measure then assumes the particularly elegant interpretation as the harmonic mean of the distances to individual clusters[4].

Regression-Wise Clustering and Correction Factors

The above discussion demonstrates how the intelligent *K*-means cluster algorithm can be applied to model a QSAR model's applicability domain in descriptor space, from which a distance-to-domain measure may be constructed. However, it is also possible to use the resulting clustering to generate predictive models.

We have recently developed techniques for enhancing descriptor-based QSAR models by using intelligent *K*-means clustering in combined descriptor-activity space. The *K*-means algorithm is modified so that clusters, instead of tending to be spherical (as tends to occur with 'straight' *K*-means), are sheared in descriptor-activity space so as to lie parallel to the hyperplane of the global linear least-squares regression.

The resulting clustering, as well as modelling the shape of the dataset in descriptor space, *simultaneously* tends to group together structures with similar discrepancies from the global linear regression. For a specific cluster, the linear regression model may therefore be improved (within that cluster) by adding a 'correction factor' determined from the activity component of the cluster's centroid.

These cluster-specific 'corrected' models can be assembled into a cluster-based model over the whole domain by using 'fuzzy weightings' as before. The original global linear regression is adjusted by adding a 'composited correction factor' comprising a weighted average of the cluster-specific correction factors. The resulting model assumes the following form, with model coefficients a_j (for descriptor j) and correction factors c_k (for cluster k):

$$y = a_1x_1 + \ldots + a_Mx_M + constant + c_1z_1 + \ldots + c_Kz_K + error \qquad (2)$$

Cluster-Based Modelling

The method described above of linear least-squares regression modelling using 'correction factors' turns out to be equivalent to performing the regression on an expanded descriptor set. As can be seen from the formulation in equation (2), the additional descriptors are none other than the fuzzy weightings $z_k(x)$ given by equation (1).

These additional descriptors z_k, which are of course entirely (but non-linearly) dependent on the original descriptors, capture affinity to a particular cluster k. Although linear regression usually models global trends, the inclusion of these 'cluster affinity' descriptors allows the regression to take into account localised phenomena: *fluctuations in activity that are confined to the vicinity of a particular cluster rather than varying directly with a particular descriptor*.

The advantage of this viewpoint is that the additional 'cluster affinity' (z_k) descriptors are available to be added to any descriptor-based QSAR modelling method, not just linear least squares. For example, non-linear support vector regression[6] can operate on this expanded descriptor set. Even descriptor-based binary classification methods, such as linear discriminant analysis and support vector machine classification[6], can benefit from the localised modelling power of these additional 'cluster affinity' descriptors, which are based in this case on the 'standard' *K*-means clustering in descriptor space only.

Multiple Endpoints

One key observation of this 'correction factors' method is that the clustering was simultaneously modelling the dataset's lie in descriptor space *and* in response (activity) space. The effect of this was to ensure that the clusters are oriented towards amenability to modelling with cluster-based correction terms or additional 'cluster affinity' descriptors.

This concept of simultaneously modelling inputs and responses has echoes of the 'counter-propagation neural network'[7]. The central layer of 'hidden nodes' in this two-layer neural network topology can be viewed as simultaneously mapping the inputs and responses using a Kohonen self-organising map. This symmetric topology results in a network that can be used as a prediction tool that may be run in either direction: predicting responses from inputs, or generating inputs that are likely to lead to a given response.

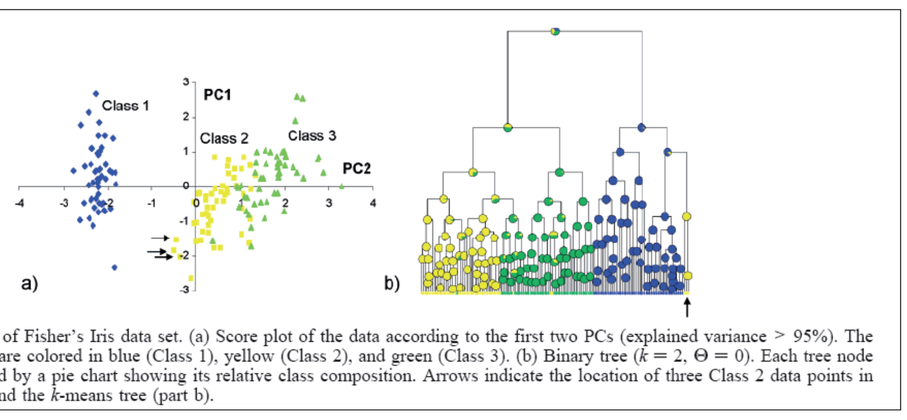
With these observations in mind, we seek an analogous method for applying the intelligent *K*-means algorithm to 'multiple endpoint' QSAR/QSPR in which the response space consists of a number of binary properties.

Once such a method has been found, it can be applied to a 'multiple activity' QSAR scenario where a response comprises several *quantitative* activities. Each quantitative activity *A* is quantised into a binary property {*A is ACTIVE*} by specifying a threshold activity value that partitions the range of activities into two classes: *ACTIVE* and *INACTIVE*. Alternatively, if both extremes of the range of activity values are of independent significance, then two binary properties, {*A is LOW*} and {*A is HIGH*}, may be defined by specifying two thresholds that partition the activity range into three classes: *LOW*, *INTERMEDIATE*, and *HIGH*. A scenario involving a mixture of binary properties and quantitative activities may also be converted to a multiple property situation in this way, as described already by many authors[8-10, for example].

Clustering in a Multiple Endpoint Environment

Let us consider first clustering in response space only. A multiple property response for a structure is represented as a sequence of bits – a 'biological fingerprint' – with a 'one' bit denoting expression of that property in the structure. The structures' biological fingerprints can then be compared using measures such as 'Tanimoto' similarity or hamming distance. This allows them to be clustered using any of the numerous well-documented methods available for similarity-based clustering.

Of particular interest are the so-called 'hierarchical agglomerative' algorithms for similarity-based clustering. This class of algorithm involves beginning with each structure in its own distinct 'singleton' cluster. The clustering then proceeds by merging clusters – two at a time chosen according to some cluster similarity criterion – until only one super-cluster (containing all structures) remains. This is very similar to the approach taken by other authors[10] using the *K*-means algorithm.



Clustering of Fisher's Iris data set. (a) Score plot of the data according to the first two PCs (explained variance > 95%). The different data classes are colored in blue (Class 1), yellow (Class 2), and green (Class 3). (b) Binary tree ($k = 2$, $\Theta = 0$). Each tree node (cluster) is represented by a pie chart showing its relative class composition. Arrows indicate the location of three Class 2 data points in the PC plot (part a) and the *k*-means tree (part b).

Figure 3. *K*-means clustering of Fisher's Iris data set[10]

Each level of the resulting hierarchy depicts a different clustering lying in the spectrum from high sensitivity (many clusters) to low sensitivity (few clusters).

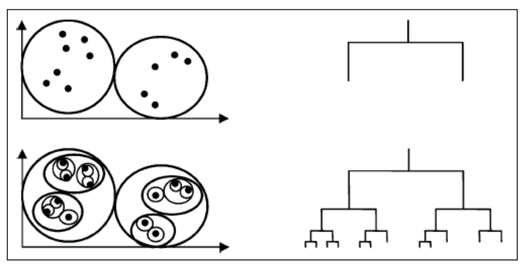


Figure 4. Example of hierarchical clustering[10]

This hierarchy may be presented visually on a graph as a binary tree or *dendogram*, with structures on the x-axis (ordered such that every cluster consists of contiguous structures). The y-axis represents dissimilarity, and each node representing the merger of a pair of clusters is plotted with a y-value according to the (dis)similarity of the clusters being merged.

A particular clustering can then be selected, either according to a desired number of clusters, or by taking a horizontal 'slice' through the dendogram intersecting the y-axis at a specified threshold value of (dis)similarity.

Guided Clustering

We return to the goal of simultaneously clustering in descriptor space and in response space. One crude way of achieving this is to begin by clustering in response space only, resulting in 'biological' clusters – biological fingerprints. Intelligent *K*-means clustering is then independently performed on descriptor space only, yielding 'chemical' clusters. A contingency table of these two cluster sets indicates which chemical clusters correspond to (i.e. have a significant number of structures in common with) which biological clusters, and vice versa [8, 11].

	FPC1	FPC2	FPC3	FPC4	FPC5	FPC6	FPC7	FPC8	FPC9	FPC10	FPC11	FPC12	FPC13	FPC14	FPC15	FPC16	FPC17	FPC18
DC1	10%	0%	2%	5%	50%	0.50%	1.50%	7%	0%	0.50%	0.50%	3%	15%	1%	1%	2%	0.50%	0.50%
DC2	0.50%	1%	85%	2.50%	1%	0.50%	0.50%	0%	1%	0.50%	1.50%	0%	5%	0%	0.50%	0%	0.50%	0%
DC3	16%	1%	0.50%	17.50%	0%	0.50%	1.50%	15%	0.50%	2.50%	18%	2%	0.50%	0.50%	4%	20%	0%	0%
DC4	5%	0%	17%	0.50%	0.50%	0%	20%	0.50%	1.50%	22%	1%	2%	0.50%	0.50%	19%	5%	5%	0%
DC5	1%	3%	1%	0.50%	0.50%	26%	0.50%	1.50%	0.50%	0.50%	30%	1%	3%	1%	25%	1%	2%	1%
DC6	90%	0%	1%	0.50%	0%	0%	0.50%	1%	0.50%	0.50%	0%	0.50%	0.50%	1%	1%	0.50%	0.50%	2%
DC7	0.50%	2%	1%	2.50%	88%	0.50%	0.50%	0%	1%	0%	1%	0.50%	0.50%	0%	1%	0.50%	0%	0.50%
DC8	0%	0%	3%	19%	4%	0.50%	0.50%	0.50%	21%	3%	0.50%	32%	3%	1%	1%	1%	4%	5%
DC9	2%	31%	0%	1%	16%	17%	1%	23%	0%	1%	0.50%	0.50%	2%	4%	0%	0%	1%	0%
DC10	0.50%	0.50%	1%	4%	2%	3%	87%	0%	0%	0%	0.50%	0%	0.50%	0%	0%	0%	0.50%	0.50%
DC11	1%	0%	0%	0%	0%	0%	0%	2%	3%	0%	0%	4%	85%	0%	1%	1%	0.50%	0.50%
DC12	21%	16%	9%	1%	0.50%	0%	0.50%	0%	1%	1%	1%	2%	1%	24%	0%	0%	18%	3%
DC13	42%	31	0%	2%	3%	0.50%	0.50%	0%	16%	0%	0.50%	0.50%	0.50%	0.50%	2%	1%	0%	0%
DC14	0%	0.50%	0%	0%	0.50%	0%	0.50%	0.50%	0%	0%	81%	0.50%	0%	0%	0.50%	16%	0%	0%
DC15	1.50%	7%	0%	0.50%	0.50%	3%	15%	1%	1%	2%	0.50%	20%	16%	18%	0%	3%	7%	2%
DC16	0.50%	0%	1%	0.50%	1.50%	0%	5%	0%	0.50%	0%	0.50%	0.50%	31%	4%	17%	21%	2%	16%
DC17	1.50%	16%	0.50%	2.50%	18%	2%	0.50%	0.50%	4%	20%	0%	0.50%	1%	0%	17%	0%	16%	0%
DC18	20%	0.50%	1.50%	22%	1%	2%	0.50%	0.50%	19%	5%	5%	1%	0%	0%	1%	1%	0.50%	17%
DC19	0.50%	1.50%	0.50%	0.50%	30%	1%	3%	1%	25%	1%	2%	23%	1%	16%	2%	1%	1%	0%
DC20	0.50%	1%	0.50%	0.50%	0%	0.50%	0.50%	1%	1%	0.50%	0.50%	0.50%	90%	0.50%	0.50%	0.50%	0.50%	0%
DC21	0.50%	0%	1%	0%	1%	0.50%	0.50%	0%	1%	0.50%	0%	87%	4%	0.50%	0.50%	3%	0%	0%
DC22	0.50%	0.50%	21%	3%	0.50%	32%	3%	1%	1%	1%	4%	21%	0%	0.50%	5%	3%	1%	1%
DC23	1%	23%	0%	1%	0.50%	0.50%	2%	4%	0%	0%	1%	4%	3%	31%	4%	24%	0.50%	0.50%
DC24	87%	0%	0%	0%	0.50%	0%	0.50%	0%	0%	0%	0.50%	0%	0.50%	4%	2%	1%	0%	2%
DC25	0%	2%	3%	0%	0%	4%	85%	0%	1%	1%	0.50%	0.50%	0%	1%	0%	0%	2%	0%
DC26	0.50%	0%	1%	1%	1%	2%	1%	24%	0%	0%	18%	0%	0.50%	31%	16%	1%	1%	2%
DC27	0.50%	0%	16%	0%	0.50%	0.50%	0.50%	0.50%	2%	1%	0%	0.50%	40%	1%	4%	24%	8%	1%
DC28	0.50%	0.50%	0%	0%	81%	0.50%	0%	0%	0.50%	16%	0%	0%	0%	0%	1%	0%	0%	0%
DC29	0.50%	3%	15%	1%	1%	2%	0.50%	0.50%	0%	0%	10%	0%	2%	5%	50%	0.50%	1.50%	7%
DC30	1.50%	0%	5%	0%	0.50%	0%	0.50%	0%	0%	0%	0.50%	1%	85%	2.50%	1%	0.50%	0.50%	0%
DC31	18%	2%	0.50%	0.50%	4%	20%	0%	0%	0%	0%	16%	1%	0.50%	17.50%	0%	0.50%	1.50%	15%
DC32	1%	2%	0.50%	0.50%	19%	5%	5%	0%	0%	0%	5%	0%	17%	0.50%	0.50%	0%	20%	0.50%
DC33	30%	1%	3%	1%	25%	1%	2%	1%	0%	1%	1%	3%	1%	0.50%	0.50%	26%	0.50%	1.50%
DC34	0%	0.50%	0.50%	1%	1%	0.50%	0.50%	2%	0%	0%	90%	0%	1%	0.50%	0%	0%	0.50%	1%
DC35	1%	0.50%	0.50%	0%	1%	0.50%	0%	0.50%	0%	0%	0.50%	2%	1%	2.50%	88%	0.50%	0.50%	0%
DC36	0.50%	32%	3%	1%	1%	1%	4%	5%	0%	1%	0%	0%	3%	19%	4%	0.50%	0.50%	0.50%
DC37	0.50%	0.50%	2%	4%	0%	0%	1%	0%	0%	0%	2%	31%	0%	1%	16%	17%	1%	23%
DC38	0.50%	0%	0.50%	0%	0%	0%	0.50%	0.50%	0%	0%	0.50%	0.50%	1%	4%	2%	3%	87%	0%
DC39	0%	4%	85%	0%	1%	1%	0.50%	0.50%	2%	0%	1%	0%	0%	0%	0%	0%	0%	2%

Figure 5. PredictionBase: A contingency table between 'chemical' and 'biological' clusters

The numbers in Figure 5 above represent the percentage of the compounds overlap between the clusters, and the colour scheme corresponds to the overlap threshold – Low (0 to 14.9), Intermediate (15 to 79.9), High (80 to 100).

An alternative, more advanced method also starts with a few (less than 10) clusters in response space only. For each biological cluster, a measure of similarity (of a test structure) to that cluster can yield a quantitative 'indicator' of the biological property combinations represented by that cluster. These biological indicators can then be used (in place of activity) to guide the intelligent *K*-means algorithm exactly as for the regression-wise clustering described earlier.

In a simultaneous clustering in response and descriptor spaces, whether achieved by one of these methods or by a different approach, the clusters represent structures that are both chemically related and biologically related. Such a clustering can be used in two ways.

Firstly, given a new structure (with known descriptor values), one can determine its cluster in descriptor space. The chemical-biological correspondence in the clustering indicates its predicted cluster (or probable clusters) in response space (biological fingerprints). Secondly, given a biological cluster of particular interest, the chemical-biological correspondence of the clustering may be used in the opposite direction to highlight clusters in descriptor space that are likely to contain structures with the desired biological properties.

References

[1] http://www.cerep.com

[2] http://www.genedata.com

[3] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC, London, 2005.

[4] R. Stanforth, E. Kolossov, B. Mirkin, A Measure of Domain of Applicability for QSAR Modelling Based on Intelligent K-Means Clustering, in press, 2006.

[5] J. C. Bezdek, S. K. Pal (Eds.), Fuzzy Models for Pattern Recognition, IEEE Press, New York, 1992.

[6] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[7] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley, New York, 1999.

[8] J. S. Melnick et al., An efficient rapid system for profiling the cellular activities of molecular libraries, PNAS, 2006, 103(9), 3153 – 3158; available online at http://www.pnas.org/cgi/doi/10.1073/pnas.0511292103

[9] J. Klekota, E. Brauner, F. P. Roth, and S. L. Schreiber, Using High-Throughput Screening Data To Discriminate Compounds with Single-Target Effects from Those with Side Effects, J. Chem. Inf. Model., 2006, 46, 1549 – 1562

[10] A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup, and G. Schneider, A Hierarchical Clustering Approach for Large Compound Libraries, J. Chem. Inf. Model., 2005, 45, 807 – 815

[11] J. W. Godden, Ling Xue, F. L. Stahura, and J. Bajorath, Searching for molecules with similar biological activity: analysis by fingerprint profiling, Pacific Symposium on Biocomputing, 2000, 5, 563-572