CASE A: CLUSTERING FOR SAR OR MECHANISM STUDIES

When performing SAR or mechanism of action studies, it is important to use clustering methods that tend to produce clusters that are pure with regard to the mode of action and possibly activity. Here we analyzed several data sets using only structural information (non-supervised clustering).

For comparison purposes, we have included two traditional methods: Jarvis-Patrick and k-modes. We have shown with each example the single coverage parameter for scaffolds. In each case, we adjusted the parameters for Jarvis-Patrick and k-modes to give approximately the same number of clusters and coverage of compounds in clusters as scaffold-directed clustering gave. This is a tremendous help for selecting the parameters, since otherwise there is no basis but trial and error for selecting the clustering parameters. Thus, the results for these two methods are partially optimized with the knowledge of the approximate number of structural clusters present in the various data sets.

This example shows that the clusters generated by scaffolds have a higher purity of mechanism than those generated by either JP or k-modes, even when the latter use the results of scaffold-directed clustering.

In this example, we tried to push the performance of unsupervised clustering to see if it would be possible to generate clusters that were pure with respect to activity. The quinolone compounds in this data set are classified into clusters that were purely active or inactive compounds using scaffold-directed clustering with the coverage parameter increased from the default. By comparing, only 37% of the JP clusters were pure, and again, this result is based on knowing the results of scaffold. For unexplained reasons, use of k-modes on this data set (with parameters set based on scaffold-directed clustering) gave a very high proportion of clusters that were inactive.

CASE B: PREFERRED SCAFFOLDS

The new scaffold method can be used to identify preferred scaffolds. For this example, we selected a chemical library that was assembled to provide compounds for screening that would be likely to, or that have been shown to be, active against a specific protein target, oxo-oxygenase-lamotrigine/acylase. The scaffolds generated by scaffold-directed clustering represent an abstract way, the general families of compounds known to be active (or likely to be active) against this family of enzymes. This library is a product of Olava (Kyiv, Ukraine). We selected a second set comprised of all of the analgesics in the WDI database. The scaffolds as SMARTS queries can be applied against another library, corporate or vendor, to identify compounds that may also be active.

The data reduction from compounds to scaffolds varied from 10:1 to 25:1. The amount of data reduction is a feature of the compound sets and could not be easily predicted or estimated ahead of time. This shows the difficulty encountered when traditional clustering methods where parameters are selected based on the number of clusters or the number of singlets.

Examples of the scaffolds highlighted on the smallest molecule of a cluster are shown below.