PickR: Pick diverse R-groups for library design using 3D electrostatics and shape

P. Tosco, S. Sciammetta, T. Cheeseright, M. Mackey

Cresset, Cambridgeshire, UK paolo@cresset-group.com cresset-group.com

Introduction

Diverse library design and enumeration is an important technique for generating new chemical matter for hit or lead finding. The ultimate goal is to obtain the broadest coverage of chemical space while minimizing the number of molecules to buy or synthesize.

Existing techniques generally use 2D methods involving chemical features or structural graphs to assess the similarity of any two compounds. Structures which appear to largely differ in functional group decoration may give rise to quite similar steric and electrostatic properties.

In this contribution we show how 3D electrostatic and shape similarity can bring a much richer, more realistic description of molecular interactions. However, it introduces conformational sampling into the problem, significantly increasing the size and complexity of the calculations.

Method

The PickR[1] algorithm utilizes the concept that most libraries are constructed using a combinatorial paradigm, such that the selection of the final molecules to be included in the library can be simplified to the selection of a suitable range of building blocks, or R-groups. To assess the diversity of these R-groups, we align all reagents on a common bond, usually the bond formed in the combinatorial reaction, and compute the electrostatic[2] and shape similarity of every pair of conformations. As the alignment along a bond involves a rotational degree of freedom, we sample multiple mutual arrangements of each reagent pair to make sure that the best steric and electrostatic overlap is attained (Figure 1).

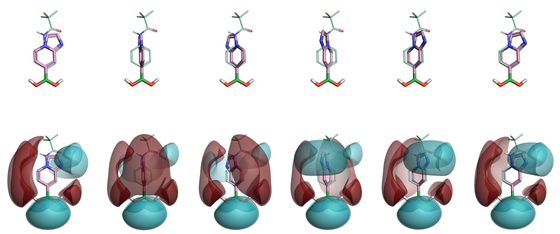

Figure 1. Alignment and rotation around the B-C bond and the generated electrostatic map.

This procedure leads to a single similarity value for each reagent pair, which is collected into a similarity matrix.

Clustering

Clustering (*k*-medoids) of the similarity matrix yields to a diverse pick of R-groups which are prioritized for inclusion in the library to be synthesized or acquired from a vendor.

Application to amino acid side chain selection

A dataset of approximately 1,000 amino acid raw reagents from eMolecules were processed to convert the side chains into R-groups, while replacing the C-alpha atom with iodine (all other iodine-containing reagents were excluded as were those containing Br and those with side chains >150Da). Using PickR, a 3D similarity matrix for the side chains was generated, aligning on the C-alpha to C-beta bond. 100 clusters were requested initially (Figure 2).

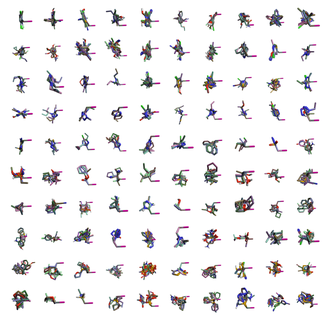

Figure 2. 3D representations of all 100 clusters generated from amino acid side chains, aligned to each other using the I-C bond of the fragments.

Looking at the results, there are some very nice relationships. For example, in cluster 2, together with tyrosine, are other phenolic side chains but also an indazole that contains the donor-acceptor motif (Figure 3).

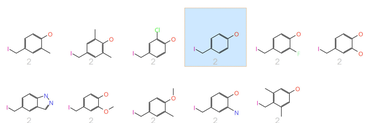

Figure 3. 2D representations of all the side chains in the same cluster as tyrosine (highlighted).

The cluster containing the phenylalanine side chain highlights the major difference of PickR over other methods – R-groups are clustered on 3D electrostatic properties. Hence, together with the phenylalanine side chain there are thiophenes and pyrroles but few other aromatics – pyridine and pyrimidines go to their own clusters because using electrostatics they are quite different to a plain phenyl ring (Figure 4).

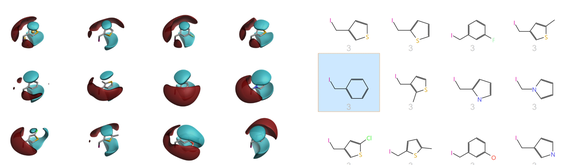

Figure 4. 3D and 2D pictures of all side chains in the phenylalanine cluster.

Application to 200 boronic acids

200 commercially available boronic acids were randomly selected from eMolecules and submitted to PickR to generate 20 clusters (Figure 5).

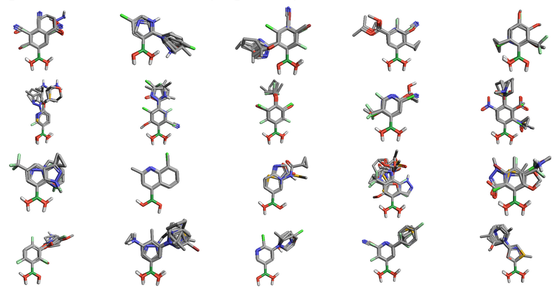

Figure 5. 3D representations of all 20 clusters generated from 200 boronic acids, aligned to each other using the B-C bond of the boronic acids.

Comparison to 2D methods

Using a 3D similarity metric generates significantly different results to established 2D fingerprint-based methods (Figure 6). In practice, we expect most practitioners to use a combination of both 3D and 2D similarity matrices for final building block selection as each has value.

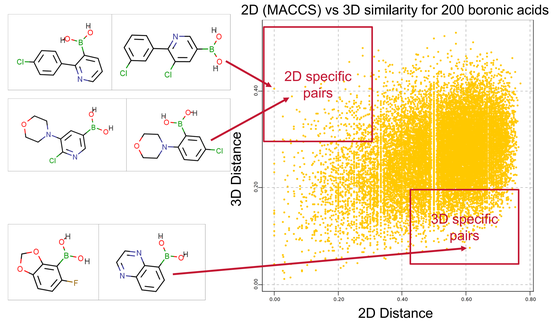

Figure 6. Comparison of pairwise 2D vs 3D similarities.

Calculation time

The need to consider conformations increases the calculation time considerably over 2D methods. However, the application has been engineered to automatically split and submit the calculations to a queuing system or to the local workstation. With a moderately sized compute cluster, datasets of 10,000 reagents are easily accessible in a reasonable time (Figure 7).

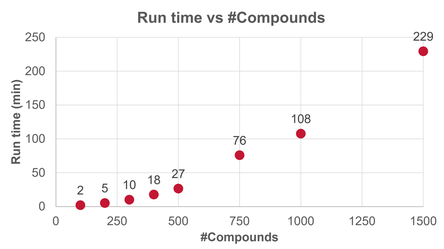

Figure 7. Calculation time *vs* dataset size on a single workstation.

Application to 5,500 boronic acids

A set of 5,500 commercially available boronic acids was selected from the eMolecules database using atom count and rotatable bond limits (#atoms<18; each rotatable bond counts as 2 atoms). PickR was applied using the default options as follows:

```
pickr -s '[B:1][#6:2]' \    #bond to break
      -Q sge \              #use SGE queueing
      -j 200 \              #split to 200 jobs
      -v \                  #verbose output
      boronics.smi          #mols (sdf also)
```

The calculation took 12 hours on 150 CPU cores and generated 550 clusters (10% of the dataset size, the default).

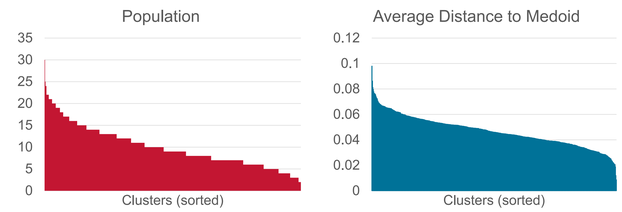

Figure 8. Distribution of cluster populations and average distance of cluster items to their medoid.

The largest cluster contained 30 compounds, the smallest two compounds. The average distance to a cluster medoid was 0.05 (Figure 8).

The utility of the method is neatly demonstrated by the populations of the top two clusters which nicely separate 3-substituted phenyls from 3-substituted 4-pyridyls (shown in Figure 9 with distance to medoid).

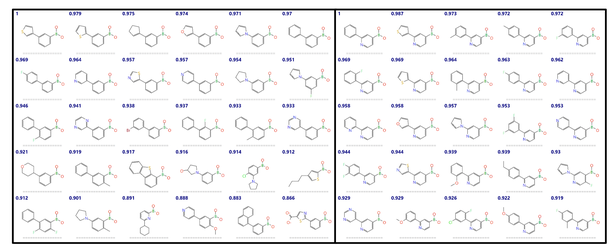

Figure 9. The two largest clusters of boronic acids with the distance to the medoid.

Conclusion

PickR is an excellent method for clustering reagents for library design. The method enables consideration of conformational and electrostatic effects giving a more diverse design of the reagent library. Although computationally expensive, it can be applied to datasets of several thousand reagents easily using the built-in job distribution options.

References

1. http://www.cresset-group.com/pickr
2. Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model*. **2006**, 46 (2), 665-676

© Cresset®