A MEDICINAL CHEMISTRY BASED MEASURE OF R GROUP SIMILARITY

Noel M. O'Boyle, Roger A. Sayle

NextMove Software Ltd, Cambridge, UK

Motivation

Medicinal chemistry projects often proceed by replacement of R groups with others that are in some way similar. However, commonly used methods to measure molecular similarity (e.g. ECFP4 fingerprints) perform poorly when comparing R groups. Previous approaches include descriptor-based measures of R group similarity [1,2] and adaptation of fingerprints to R groups [3].

Here we describe the use of medicinal chemistry project data itself (as present in ChEMBL and patents) to suggest R groups that are similar to a query.

ChEMBL dataset

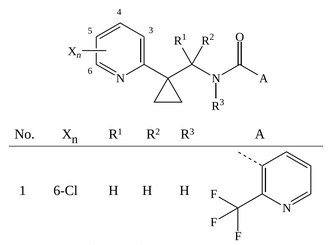
IC_{50}, EC_{50} and K_i data (marked as non-duplicate) were exported from ChEMBL25. Molecules from the same document (either a publication or a patent) were fragmented according to the procedure described in [4] and matched series identified. By summing over all documents, frequencies of co-occurrence of particular R groups in a series were calculated. Note that even where a particular co-occurrence occurred multiple times in a document, it only counted as a single value towards the total.

US patent dataset

TABLE 9

Inhibition of CETP Activity by Examples in Reconstituted Buffer Assay.

Ex. No.	IC_{50} (μM)	Ex. No.	IC_{50} (μM)	Ex. No.	IC_{50} (μM)
249	0.020	419	0.19	425	0.34
244	0.029	230	0.20	514	0.34
634	0.032	248	0.20	237	0.35
221	0.034	266	0.20	399	0.35
229	0.034	378	0.20	645	0.35

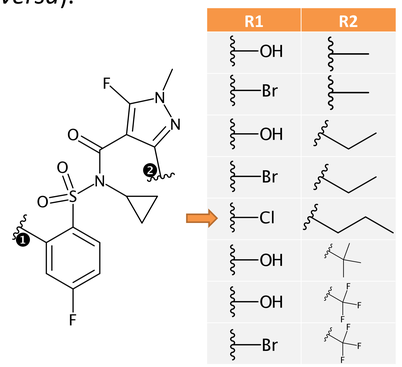


LeadMine was used to extract molecules associated with bioactivity (as on the left, from US20030100559A1) or present in Markush tables (right, from US20160309717A1).

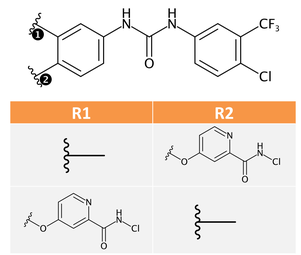
NextMove Software's chemical text-mining tool LeadMine was used to extract 2.9M molecules from pharmaceutically-related USPTO patent applications. To ensure independent observation of co-occurrences, patent applications were grouped into chemically-related patent families (CRPF) [5]. Molecules from the same CRPF were fragmented and analysed as for the ChEMBL dataset.

Infer R group replacement from double-cut data

The datasets available from ChEMBL and patents typically include only a portion of the full project data. To maximise recovery of R group co-occurrences, additional co-occurrences were inferred from double-cut matched series. Where R1 and R2 positions are related by symmetry, only cases where the R1 group was the same as R2 were retained. To minimise spurious co-occurrences due to transposition of R groups, series were discarded if any R1 group also appeared at R2 (or *vice versa*).



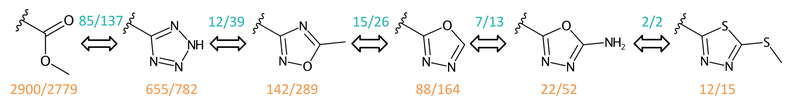
In this matched series from US patent application US20090137611A1, the correspondence between propyl and other R2 groups is missed by single-cut analysis. Similarly for chloride and other R1 groups.



This series from US patent application US20090068146A1 was discarded as described above.

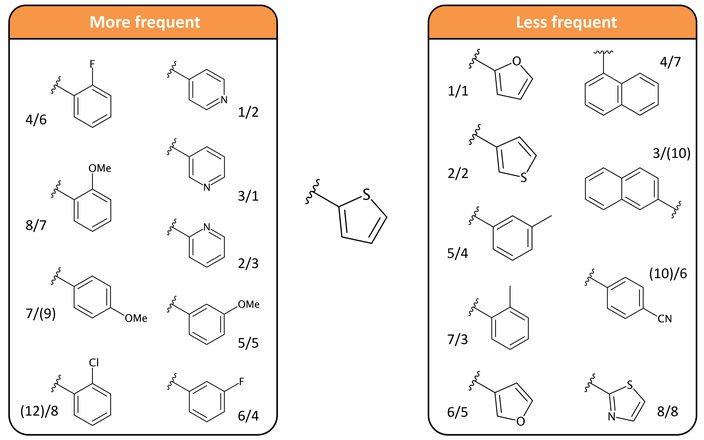
Identify similar R groups

Given a query R group, we use the frequency of occurrence of R groups as components of a matched series to divide the remaining R groups into those that would typically be synthesised before the query (and are therefore more frequent) and those that would only be synthesised after (less frequent). For the latter, the most similar are those with the highest co-occurrence. In the former case, a balance must be struck between higher co-occurrence and lower frequency; we use co-occurrence divided by frequency to identify similar R groups.



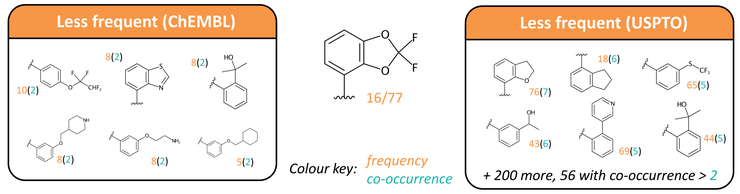
Examples of frequency (orange) and co-occurrence (blue) between R groups in ChEMBL/USPTO matched series

Example: The 8 R groups most similar to thiophen-2-yl



The 8 most similar R groups among those more frequent are shown on the left, while similar but less frequent R groups are shown on the right. The numbers indicate the rank according to the ChEMBL database/USPTO dataset.

Comparison of dataset results for an uncommon R group



Conclusions

Despite the differences between the two datasets, the rankings from ChEMBL and US patents are often in very good agreement and appear reasonable. However, for less frequent R groups, the patent dataset provides more suggestions than the ChEMBL dataset due to the larger table size in a typical pharmaceutical patent.

References

[1] P. Ertl. *J. Mol. Graph. Model.* **1998**, *16*, 11.
[2] J.D. Holliday, S.P. Jelfs, P. Willett, P. Gedeck. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, **2**, 406.
[3] S. Tamura, T. Miyao, K. Funatsu. *J. Chem. Inf. Model.* **2019**, In press.
[4] N.M. O'Boyle, J. Boström, R.A. Sayle, A. Gill. *J. Med. Chem.* **2014**, *57*, 2704.
[5] https://nextmovesoftware.com/blog/2017/07/04/chemically-related-patent-families/