BEYOND BINARY: SCALED MOLECULAR FINGERPRINTS FOR MAXIMUM DIVERSITY PICKING

Vladimir Berishvili, Alexei Pushechnikov

ChemDiv, 12760 High Bluff Drive, Ste. 370 San Diego, CA 92130, USA

E-mail: vberishvili@chemdiv.com

INTRODUCTION

In the practice of a medicinal chemist choosing the most diverse subset of chemical compounds is a rather common task. This task occurs most often while conducting virtual high-throughput screening for selecting compounds to probe a novel biological target. Given a limited budget, selecting a diverse subset of compounds from a larger library can maximize the chemical space coverage, thereby improving chances of finding good starting points for a discovery program.

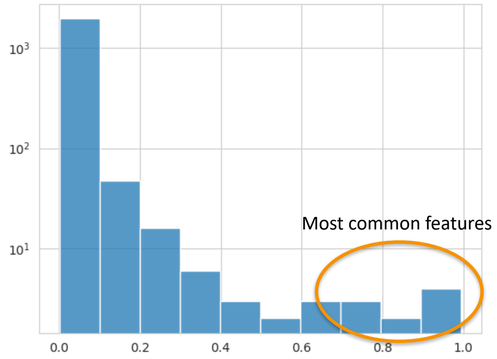
The diversity picking involves three key steps: (1) the choice of molecular representation, (2) the choice of either similarity or distance metric, and (3) the choice of an algorithm that leads to a global (or local) optimum for a given metric of diversity. We focus in this report on the first step.

A widely used combination of binary fingerprints (e.g. ECFP, MACCS [1,2]) with either Tversky [3] or cosine similarities advocates for an absolute equality of all bits (molecular features) in a representation. Despite the simplicity of this approach, it is probably not the most efficient one for a task of a chemical diversity selection, given the huge differences in the frequencies of different features.

BACKGROUND

Not all fingerprint-encoded features are equally common, so they carry different amounts of information about molecular structure in the context of a library.

Count of ECFP fingerprint bits (ChEMBL database)



Percent of ChEMBL compounds where particular bit (feature) appears

MATERIAL & METHODS

Fingerprints Scaling

$$Ibf(b,L) = log(\frac{|L|+1}{1+|\{s \in L : b \in s\}|}),$$

$$FPscaled = FP \cdot Ibf(L),$$

where $Ibf(b,L)$ – inverse bit frequency for library L,

FP – binary fingerprint (ECFP 1024 bits),

$FPscaled$ – scaled fingerprint,

L – compound library,

b – feature (bit) from the FP vector,

$|\{s \in L : b \in s\}|$ – number of compounds where feature b appears

Diversity Assessment

$$Diversity = \frac{100\,\%}{2|L|^2} \sum_{i,j} Dist(i,j),$$

where $Dist(i,j)$ – cosine distance between compounds i and j

Databases used to learn scaled fingerprints

- ChEMBL 27 — chemical space of biologically active compounds
- ChemDiv Stock (~1.6M compounds)

RESULTS: DIVERSITY

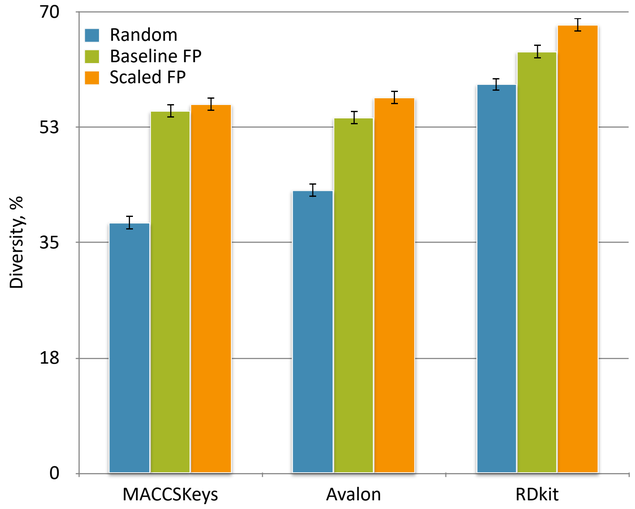
Diversity picking was performed using a popular MinMax algorithm using standard and scaled fingerprints. For comparison of diversities of the libraries obtained by different methods, we used several unrelated standard fingerprints. Also, to use the same algorithm for both the baseline and scaled versions of the fingerprint one have to convert scaled version back to the binary format. However, after scaling one can choose a cutoff value based on the calculated feature importance. The underling rationale is that the diversity of the most common structural fragments is more important for the overall dataset diversity than the diversity of the relatively rare features.

The proposed algorithm is shown below.

Algorithm 1 Diversity Picking

```
procedure DIVERSITY–PICKING
    Scale fingerprint
    Select cutoff value ▷ Select based on distribution of feature importance
    Truncate scaled fingerprint
    Use MinMax algorithm
end procedure
procedure DIVERSITY–EVALUATION
    for each fingerprint i ∈ FPs do        ▷ Set of standard fingerprints
        Calculate diversity using i
    end for
end procedure
```

Average diversity of the selected 1K compounds from the ChemDiv inventory calculated using standard fingerprints

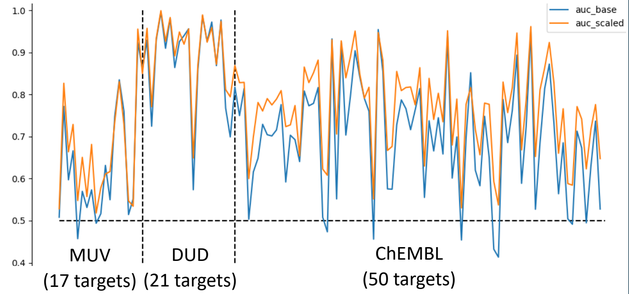


ECFP 1024(RDKit implementation) was used as a baseline FP

RESULTS: VIRTUAL SCREENING

Using the benchmarking platform [4], the performance of ECFP 1024 bits fingerprint and its scaled variant was assessed over 88 targets. Test molecules are ranked based on the cosine similarity to the actives in the training set. Only the highest similarity value is considered for each test molecule.

Average performance of fingerprints measured with AUC



CONCLUSIONS

Using bit frequency normalization (akin to the popular natural language processing algorithm TF-IDF), one could obtain not only the representation of a molecular structure as such, but also the representation of a specific molecular structure in the context of a chemical compound library. Thus, the resulting floating-point vector encodes the information about relative frequencies of various molecular features across an entire library. Also we have demonstrated that the proposed normalization can be a quite useful addition to a virtual screening workflow.

We have shown that using this approach one can get:

- more diverse compound library compared to results obtained via standard fingerprints
- higher AUC ROC values for the similarity based virtual screening as shown for the popular fingerprints benchmarks [4]

REFERENCES

1. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model., 50 (5): 742-754, 2010.
2. Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of MDL keys for use in drug discovery. Journal of chemical information and computer sciences, 42:1273–1280, 2002.
3. Tversky A. Features of Similarity. Psychological Review, 84 (4): 327–352, 1977.
4. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform 5, 26, 2013.