Turbo Similarity Searching

Jérôme Hert*, Peter Willett*, David J. Wilton*, Kamal Azzaoui†, Edgar Jacoby† and Ansgar Schuffenhauer†.

INTRODUCTION

The similar property principle is the basis for many chemoinformatics applications developed to support drug discovery and although there are many exceptions, its general applicability is widely acknowledge [1,2].

The principle states that if the similarity between a reference molecule and the compounds in a library is calculated, the compounds at the top of the list ranked in order of decreasing similarity score, the *nearest neighbours* (NNs), are likely to have similar properties / bioactivity to the reference structure.

The method reported herein is based on the principle and performs similarity searches using not only an initial structure but also its NNs as references; the scores of the different searches are then combined using the MAX fusion rule to produce the final ranking. We have demonstrated previously that, when a small set of compounds is available, data fusion permits a substantial increase in the recall of active compounds over conventional similarity searching (SS) [3].

We compare, here, the effectiveness of this new approach, called turbo similarity searching (TSS) to the effectiveness of SS. TSS uses only information about the NNs, assuming that they are active; if it can be demonstrated that TSS is superior to SS, we have a extremely simple way of enhancing conventional SS.

WHY TURBO?

A turbocharger uses the output of an engine, the exhaust gases, to increase the pressure in that engine, thus resulting in improved performance. The algorithm used here is called "turbo", because it uses the output of a similarity search, the NNs, to increase the effectiveness of that search. Additionally, TSS is a virtually free way of enhancing SS (it just requires some additional CPU time) in much the same way as a turbocharger is an almost free source of energy for an engine.

REFERENCES

[1] MA Johnson & GM Maggiora, Concepts and Applications of Molecular Similarity; John Wiley: New York, 1990.

[2] Willett *et al.*, J. Chem. Inf. Comput. Sci. **1998**, *38*, 983

[3] Hert *et al.*, J. Chem. Inf. Comput. Sci. **2004**, *44*, 1177

[4] Scitegic Inc. is at www.scitegic.com

[5] MDL Information Systems is at www.mdl.com.

METHODS

The strategy employed by TSS is illustrated in figure 1. An initial similarity search is conducted using the reference structure to identify its NNs. Additional similarity searches are carried out using a user-defined number of these NNs, and the final ranking of the library of compounds is produced by combining the scores, for all compounds, of all the similarity searches using the MAX fusion rule. This method hence requires, just as SS, only a single reference structure and assumes that the NNs exhibit a similar bioactive as the reference; its only parameter is the number of NNs that are considered as assumed active reference structures.

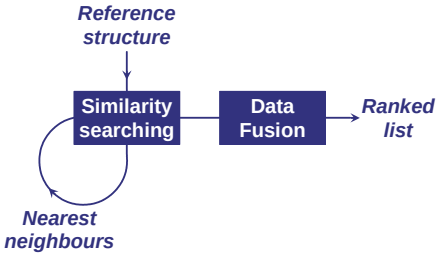

Figure 1: Illustration of the TSS algorithm.

Activity name	Nb of compounds	Self-Similarity Mean	SD
5HT3 antagonists	752	0.351	0.116
5HT1A agonists	827	0.343	0.104
5HT reuptake inhibitors	359	0.345	0.122
D2 antagonists	395	0.345	0.103
Renin inhibitors	1130	0.573	0.106
Angiotensin II AT1 antagonists	943	0.403	0.101
Thrombin inhibitors	803	0.419	0.127
Substance P antagonists	1246	0.399	0.106
HIV protease inhibitors	750	0.446	0.122
Cyclooxygenase inhibitors	636	0.268	0.093
Protein kinase C inhibitors	453	0.323	0.142

Table 1: MDDR activity classes used in the study.

EXPERIMENTAL DETAILS

The similarity scores in all of the experiments were computed using the Tanimoto coefficient, with the reference and database structures characterised by Scitegic ECFP_4 fingerprints [4].

The performance of both SS and TSS was computed by means of simulated virtual screening experiments. Specifically, we used the eleven diverse activity classes from the MDL Drug Data Report (MDDR) database [5] listed in Table 1. The database was searched by using, in turn, each of the 8,294 active compound as a reference. For each search, a note is made of how many compounds of the same activity class have been retrieved after processing 5% of the database. The results presented are the average of this recall over the 8,294 different searches.

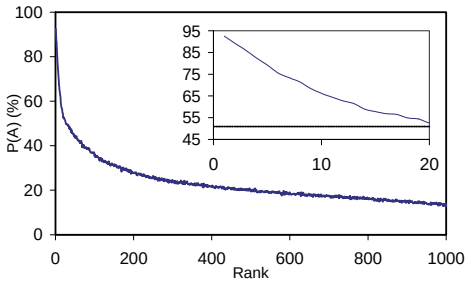

Figure 2: Average probability P(A) of a compound to be active as a function of its rank.

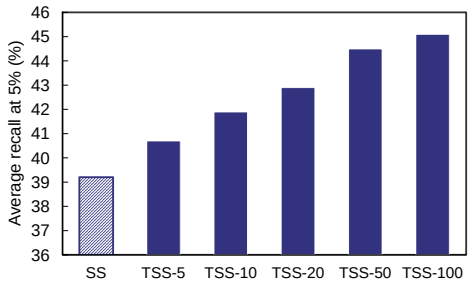

Figure 3: Comparison of the average recall at 5% of SS and TSS.

RESULTS

A initial experiment was conducted to confirm that the datasets studied here do indeed satisfy the similar property principle. The probability that a compound would be active was plotted against its rank. This probability was obtained by averaging the number of times a compound at a given rank was active. The results are show in Figure 2 where it can be seen, e.g., that the first 20 NNs have a probability of being active that is greater than 50%.

SS and TSS are compared in Figure 3 when TSS is carried out using the specified number of NNs (5, 10, 20, 50, 100) and where each bar represents the recall at 5% averaged over the 8,294 searches. It can be seen from this bar-chart that TSS always outperforms SS, even when a small number of NNs is considered. The best results are obtained with TSS-100. TSS-200 has also been evaluated but appeared to be less effective than TSS-100.

CONCLUSION

We showed that it is possible to increase the effectiveness of conventional SS by using information about the NNs resulting from the initial search. The algorithm requires no modification to existing similarity software other than the ability to fuse the outputs of multiple searches.

Our experiments involved 2D fingerprints and the Tanimoto coefficient but there seems to be no reason in principle why this approach could not be used with any other similarity measure that satisfies the similar property principle.

* *Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.*

† *Novartis Institutes for Biomedical Research, Discovery Technologies, Lead Discovery Center, Molecular and Library Informatics Program, CH-4002 Basel, Switzerland.*

ACKNOWLEDGMENT

We would like to thank the following: Novartis Institutes for Biomedical Research for funding; MDL Information Systems Inc. for the provision of the MDDR database; and Scitegic Inc. for software.