Journal of Computer-Aided Molecular Design (2005) 19: 687–692
DOI 10.1007/s10822-005-9000-8

Similarity-based descriptors (SIBAR) – A tool for safe exchange of chemical information?

Dominik Kaiser, Barbara Zdrazil & Gerhard F. Ecker*
Department of Medicinal/Pharmaceutical Chemistry, University of Vienna, Althanstrasse 14, 1090 Wien, Austria

Received 17 May 2005; accepted 27 June 2005

Key words: ADME, ChemMask, SIBAR-descriptors, similarity search, VSA-descriptors

Summary

Exchange of chemical information without disclosure of the respective structures would greatly increase the data sets available for model building. Within the framework of the ChemMask project we explored the principal applicability of SIBAR-descriptors to mask chemical structures. SIBAR is based on calculation of similarity values for each compound of the training set to a set of reference compounds. Although the SIBAR-approach *per se* does not allow to unambiguously trace back the chemical structure of a compound, similarity searching in a 1.5 million compound database spiked with compounds structurally analogous to the query structure lead to the retrieval of compounds structurally and pharmacologically highly analogous to the "hidden" query structure in all three examples investigated. Comparison to results obtained with the original descriptors used to calculate the SIBAR-values showed, that SIBAR indeed adds some fuzziness to the data matrix.

Introduction

Safe exchange of chemical information is of increasing importance in the drug discovery and development process. Especially in the field of prediction of ADMET properties the possibility of sharing data sets would greatly enhance the predictive power of the various models established. Additionally, also the chemical space covered by the combined data sets would be remarkably enhanced [1]. Sharing data might also be of special interest with respect to antitargets, such as P-glycoprotein [2] and the hERG potassium channel [3]. In both cases the data available in the literature are rather limited and in some cases even contradictory [4, 5]. Thus, several attempts have been started to promote open access initiatives and to generate public data warehouses [6, 7]. However, due to the current situation with respect to intellectual properties, Pharmaceutical Industry is not participating in these initiatives. Undoubtedly all these initiatives would highly benefit from algorithms and IT solutions allowing exchange of chemical information without disclosure of the corresponding structures.

On principle there are several ways to handle this task. One might be to develop a software package which allows encryption of sd-files and calculation of descriptors but does not allow access to the chemical structures. Another possibility is to simply exchange a set of descriptors but not the corresponding structures. Especially in the field of ADMET-predictions numerous descriptors have been developed which do not allow to unambiguously trace back the corresponding structures. These include e.g. field based descriptors such as Almond [8] and Volsurf [9], autocorrelation vectors

*To whom correspondence should be addressed. E-mail: gerhard.f.ecker@univie.ac.at

[10] as implemented in ADRIANA [11] and the VSA-descriptor set developed by Labute [12].

However, performing a similarity search in a database large enough to cover the whole drug like chemical space, might be a rather simple solution to overcome the attempts for hiding the structure. Assuming one receives a file containing 32 VSA descriptors derived from a "hidden" structure, you calculate the VSA descriptors for the set of 13.9 million drug like compounds currently available [13] and run a similarity search on basis of euclidian distances. Hits retrieved should be very close to the hidden query structure. Within this paper we systematically explored this approach and compared the results with those obtained when using our previously described SIBAR-descriptors [14].

Materials and methods

All computations were performed on a 2.4 GHz Pentium IV PC operating under Red Hat Linux. Descriptors were scaled to unit variance and euclidian distances between the query structures and the database compounds were calculated using our in house svl-scipt implemented in MOE [15]. The resulting table was sorted according to increasing distance and the top 20 hits for each query structure were retrieved.

Descriptors used

For similarity searching and calculation of the SIBAR-descriptors the following three sets of descriptors were used:

(1) A set of descriptors frequently used for prediction of bioavailability. This set is denoted as ADME-descriptors and include lipophilicity (logPo/w), molar refractivity (MR), molecular weight (MW), topological polar surface area (TPSA), number of H-bond donors (nHBD) and acceptors (nHBA) and the number of rotable bonds (nROT). All calculations were performed with the software package MOE.
(2) The 32 VSA-descriptors as described by Labute [12]. These descriptors are based on atomic contributions to the van der Waals surface area. logP MR and partial charge. They are widely applicable and were successful used to generate predictive models for boiling point, solubility and blood–brain barrier permeation. These descriptors are implemented in the Molecular Operating Environment (MOE) software from the Chemical Computing Group.
(3) 2D Autocorrelation vectors for a distance of 0–10 bonds for a set of 7 atom properties. The following atom properties were calculated using the software package ADRIANA [11]: atom-polarizability (α_i), sigma-electronegatives (χ_σ) and -charges (q_σ), pi-electronegatives (χ_π) and -charges (q_π), lone pair electronegatives (χ_{lp}) and total charges (q_{tot}). Subsequently, the 2D-autocorrelation vectors for distances from 0–10 bonds were calculated with ADRIANA, which resulted in a set of 77 descriptors, denoted as AUTOCORR-set.

Query structures

As query structures a highly rigid steroid (estriol), the benzodiazepine diazepam and the propafenone-type P-glycoprotein inhibitor GPV0005 were chosen.

Database used (+ spiking)

The database for similarity searching was established via merging the compound collections provided by SPECS [16], Enamine [17], Maybridge [18] and ChemDiv [19]. After removal of all duplicates this database included 1.513.388 compounds. To check whether this database already includes the query structures and analogs thereof, a similarity search using UNITY fingerprints [20] and a Tanimoto threshold of 0.6 was performed. As neither the query structures nor close structural analogs were found, the database was spiked with small sets of steroids, benzodiazepines and propafenones.

The SIBAR-approach

The basic underlying principle of SIBAR assumes that compounds similar to biological active ones should be also active and vice versa. SIBAR-descriptors are based on calculation of similarity

values between compounds under consideration and a reference set. This leads to a given number of similarity values equal to the number of reference compounds. This similarity array is further used as input vector for QSAR analyses (Figure 1). Both the reference set as well as the descriptors used for calculating the SIBAR-values may be tailored to the specific QSAR problem. Hitherto best results have been obtained when targeting ADME-problems and polyspecific proteins.

The calculation of the SIBAR-descriptors D is outlined as follows: (1) selection of a reference compound set; (2) calculation of a set of descriptors for both the training set (query structure and database) and the reference set; (3) calculation of the SIBAR descriptors D expressed as euclidian distances between the i reference compounds and the j compounds of interest using k molecular descriptors, X_k:

$$D(i,j) = \sqrt{\sum (X_{ik} - X_{jk})^2}$$

For the present study the references set described in our previous work was used [14]. This set comprises the 20 most diverse compounds from the SPECS compound library. Thus, for each compound a set of 20 SIBAR descriptors was obtained encoding the similarity to the 20 references compounds. These SIBAR-descriptors were used for calculating the euclidian distances between the query structures and the database compounds.

Results

Calculation of the SIBAR descriptors needs the knowledge of the reference set. Thus, assuming that the structures of the compounds of the reference set are not disclosed it should be impossible to trace back the chemical structure of the original compounds. In an attempt to elucidate the amount of fuzziness the SIBAR approach introduces to a data set we performed a series of similarity searches based on calculation of euclidian distances. Distances were calculated using three different sets of descriptors (ADME-set, AUTOCORR and VSA) and the results were compared to those obtained when additionally applying the SIBAR protocol. Figure 2 shows the compounds retrieved from a set of 1.5 million compounds when using diazepam, estriol and the P-glycoprotein inhibitor GPV0005 as query structures.

First of all it has to be noted that, if the query structure is also present in the database it is in any case found. This is obvious and inherent to the method applied for similarity searching, because the euclidian distance between identical compounds always is zero. Comparison of the structures of the compounds retrieved as most similar to the corresponding query structure shows that both the AUTOCORR and VSA descriptors in all three examples lead to identification of structurally and pharmacologically highly analogous compounds. In case of autocorrelation vectors the additional calculation step applied with the SIBAR approach did not change the results. With the VSA-descriptors only in case of diazepam a completely different compound was retrieved as those with highest similarity. Thus, the degree of fuzziness introduced via the additional computational step of calculating similarity values to a reference set is not sufficient to substantially hide chemical structures. However, when analyzing the

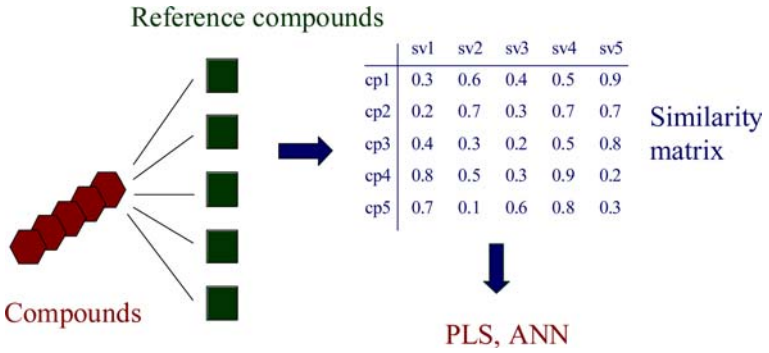

Figure 1. Workflow for the SIBAR-approach; cp: training set compound; sv: similarity value; PLS: partial least squares; ANN: artificial neural network.

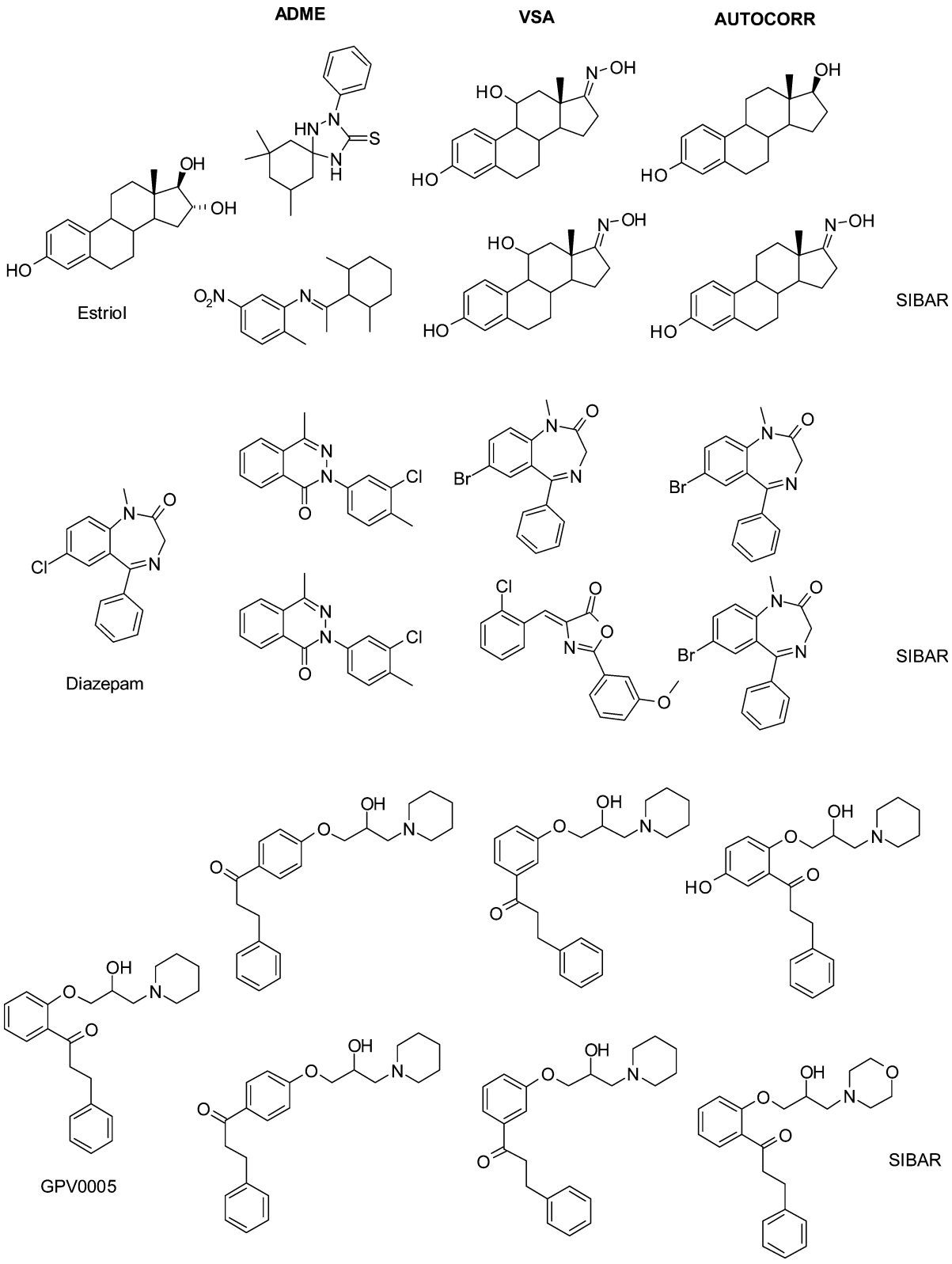
ADME
VSA
AUTOCORR
Estriol
SIBAR
Diazepam
SIBAR
GPV0005
SIBAR

Figure 2. Chemical structures of the top hits retrieved for the query structures estriol, diazepam and GPV0005 with and without applying SIBAR. Top row shows results obtained using the conventional descriptors for similarity searching, bottom row represents the hits obtained with SIBAR.

first 20 hits obtained in each run the picture is somewhat different (Table 1).

As shown in Table 1, in almost all cases the use of SIBAR gave rise to a decrease of the number of structurally analogous compounds retrieved. This is especially pronounced in case of GPV0005 and the VSA-descriptors. Using VSA alone the top 8 compounds belonged to the class of propafenones and in total 9 out of the top 20 compounds showed a propafenone scaffold. Using the VSA-descriptors to calculate 20 SIBAR-values and applying these values in the similarity search, only the closest compound showed a propafenone-type scaffold and in total 3 out of the top 20 compounds were structurally related to GPV0005. Thus, SIBAR indeed adds some fuzziness to the data set and in principle may be able to mask chemical structures.

Discussion

Safe exchange of chemical information will become increasingly important for the attempts to overcome the current bottlenecks in the drug discovery and development process. Within the framework of the ChemMask project [6] we explored the principal applicability of SIBAR-descriptors to mask chemical structures. Although the SIBAR-approach *per se* does not allow to unambiguously trace back the chemical structure of a compound, similarity searching in a 1.5 million compound database spiked with compounds structurally analogous to the query structure lead to the retrieval of compounds structurally and pharmacologically highly analogous to the "hidden" query structure. However, under real life conditions the query structure is not known and thus there is no information how close the most similar compound is. This may be overcome by expanding the hit list to the top 20 compounds, which enhances the chance of retrieving compounds with identical scaffolds. With exception of diazepam and the rather general "ADME-descriptors" this strategy was successful in all cases. Thus it seems to be mainly a matter of the size and the chemical diversity of the database how close the hits are to hidden query structure.

Although there is no absolute security that the scaffold of the query structure is amongst the top ranked hits and thus intellectual properties on principle are not affected, most companies will not be willing to take the risk and share their data. One possible solution might be to keep the reference structures secrete and allow calculation of the SIBAR-descriptors only for data sets limited in size, preferentially via a secure web-site. Subsequently companies exchange only the SIBAR-descriptors and use them for establishing ADMET-models. This ensures that similarity search is limited to small databases and the likelihood of retrieving compounds with analogous scaffolds is marginal. However, this would imply that for all problems targeted only SIBAR-descriptors can be used for data-sharing. Although they were successfully applied for P-glycoprotein inhibitors [14], this would be a rather limiting restriction for model generation.

Table 1. Number of structurally analogous compounds retrieved for the query compounds diazepam, estriol and GPV0005 (*a*/*b*).

	Estriol	Diazepam	GPV0005
ADME	0/2	0/0	2/3
ADME-SIBAR	0/0	0/0	2/3
VSA	5/8	2/2	8/9
VSA-SIBAR	1/1	0/2	1/3
AUTOCORR	7/15	1/2	7/10
AUTOCORR-SIBAR	3/8	1/1	2/4

a: Gives the number of structurally related compounds ranked on top; *b:* Gives the total number of compounds structurally related to the respective query structure; in case of estriol and VSA-descriptors (5/8) the top 5 ranked (=most similar) compounds were steroids and in total 8 steroids were amongst the 20 most similar compounds.

Acknowledgements

We gratefully acknowledge financial support from the City of Vienna, Hochschuljubilaeumsstiftung grant H-1236/2003. We are also grateful to Johnny Gasteiger for providing us with an ADRIANA license and to the Vienna University Computer Center for the UNITY software package.

References

1. Bologa, C.G., Olah, M. and Oprea, T.I., Abstracts of Papers, 229th ACS National Meeting, San Diego, CA, March 2005, pp. 13–17.
2. Gottesman, M.M., Fojo, T. and Bates, S.E., Nature Rev. Cancer, 2 (2002) 48.
3. Aronov, A.M., Drug Discov. Today, 10 (2005) 149.
4. Ekins, S., Drug Discov. Today, 9 (2004) 276.
5. Ecker, G., Chem. Today, 23 (2005) 39.
6. Allu, T.K., Bologa, C. and Oprea, T.I., Safe Exchange of Chemical Information – The ChemMask Project, http://pimento.health.unm.edu.
7. Austin, C.P., Brady, L.S., Insel, T.R. and Collins, F.S., Science, 306 (2004) 1138.
8. Pastor, M., Cruciani, G. and Watson, K.A., J. Med. Chem., 40 (1997) 4089; Almond 3.3.0, Molecular Discovery, www.moldiscovery.com.
9. Crivori, P., Cruciani, G., Carrupt, P.-A. and Testa, B., J. Med. Chem., 43 (2000) 2204; VolSurf 4.1.3, Molecular Discovery, www.moldiscovery.com.
10. Gasteiger, J., Teckentrup, A., Terfloth, L. and Spycher, S., J. Phys. Org. Chem., 16 (2003) 232.
11. ADRIANA.*Code*, Molecular Networks, Germany; www.molnet.de.
12. Labute, P., J. Mol. Graph. Mod., 18 (2000) 464.
13. ChemNavigator, www.chemnavigator.com.
14. Klein, C., Kaiser, D., Kopp, S., Chiba, P. and Ecker, G.F., J. Comp.-Aided Mol. Design, 16 (2002) 785.
15. Molecular Operating Environment, Chemical Computing Group Inc.; www.chemcomp.com.
16. SPECS Inc., www.specs.net.
17. Enamine, www.enamine.net.
18. Maybridge; www.maybridge.com.
19. ChemDiv, www.chemdiv.com.
20. UNITY database, Tripos Inc., www.tripos.com.