[6] A. König, T. Schwecke, I. Molnár, G. A. Böhm, P. A. S. Lowden, J. Staunton, P. F. Leadlay, *Eur. J. Biochem.* **1997**, *247*, 526.

[7] a) M. Oliynyk, M. J. B. Brown, J. Cortes, J. Staunton, P. F. Leadlay, *Chem. Biol.* **1996**, *3*, 833; b) R. McDaniel, A. Thamchaipenet, C. Gustafsson, H. Fu, M. Betlach, G. Ashley, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 1846.

[8] a) A. Ranganathan, M. Timoney, M. Bycroft, J. Cortes, I. P. Thomas, B. Wilkinson, L. Kellenberger, U. Hanefeld, I. S. Galloway, J. Staunton, P. F. Leadlay, *Chem. Biol.* **1999**, *6*, 731; b) C. J. Rowe, I. U. Böhm, I. P. Thomas, B. Wilkinson, B. A. Rudd, G. Foster, A. P. Blackaby, P. J. Sidebottom, Y. Roddis, A. D. Buss, J. Staunton, P. F. Leadlay, *Chem. Biol.* **2001**, *8*, 475.

[9] L. Chung, L. Liu, S. Patel, J. R. Carney, C. D. Reeves, *J. Antibiot.* **2001**, *54*, 250.

[10] P. A. S. Lowden, B. Wilkinson, G. A. Böhm, S. Handa, H. G. Floss, P. F. Leadlay, J. Staunton, *Angew. Chem.* **2001**, *113*, 799; *Angew. Chem. Int. Ed.* **2001**, *40*, 777.

[11] P. A. S. Lowden, G. A. Böhm, P. F. Leadlay, J. Staunton, *Angew. Chem.* **1996**, *108*, 3295; *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2249.

[12] R. Thiericke, J. Rohr, *Nat. Prod. Rep.* **1993**, *10*, 265.

[13] B. S. Moore, C. Hertweck, *Nat. Prod. Rep.* **2002**, *19*, 70.

[14] a) H. Nishida, T. Sakakibara, F. Aoki, T. Saito, K. Ichikawa, T. Inagaki, Y. Kojima, Y. Yamauchi, L. H. Huang, M. A. Guadliana, T. Kaneko, N. Kojima, *J. Antibiot.* **1995**, *48*, 657; b) E. I. Graziani, F. V. Ritacco, M. Y. Summers, T. M. Zabriskie, K. Yu, V. S. Bernan, M. Greenstein, G. T. Carter, *Org. Lett.* **2003**, *5*, 2385.

[15] L. E. Khaw, G. A. Böhm, S. Metcalfe, J. Staunton, P. F. Leadlay, *J. Bacteriol.* **1998**, *180*, 809.

[16] J. B. McAlpine, S. J. Swanson, M. Jackson, D. N. Whittern, *J. Antibiot.* **1991**, *44*, 688.

[17] P. Hughes, J. Musser, M. Conklin, R. Russo, *Tetrahedron Lett.* **1992**, *33*, 4739.

[18] S. Metcalfe, C. E. Canman, J. Milner, R. E. Morris, S. Goldman, M. B. Kastan, *Oncogene* **1997**, *15*, 1635.

[19] a) A. F. A. Marsden, B. Wilkinson, J. Cortes, N. J. Dunster, J. Staunton, P. F. Leadlay, *Science* **1998**, *279*, 199; b) M. S. Pacey, J. P. Dirlam, R. W. Geldart, P. F. Leadlay, H. A. I. McArthur, E. L. McCormick, R. A. Monday, T. N. O'Connell, J. Staunton, T. J. Winchester, *J. Antibiot.* **1998**, *51*, 1029.

Received: September 1, 2003 [Z 758]

Evaluation of Distance Metrics for Ligand-Based Similarity Searching

Uli Fechner and Gisbert Schneider*[a]

Ligand-based similarity metrics are frequently and successfully employed for diversity analysis and the selection of activity-enriched subsets in early-phase virtual screening and compound-library design.[1–4] As they come in many varieties, it is not trivial to choose the most appropriate concept for the task at hand. Fundamentally, these methods rely on representative reference structures (also termed "query" or "seed" structures), molecular descriptors that are correlated with biological activity, and an appropriate similarity metric. "Retrospective screening" provides a means of evaluating these factors. The basic idea is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximize the number of known actives in the subset, thereby forming a "focused library".[5] Subset selection is based on the pairwise similarity between the query structure and each molecule in the pool. The result of this calculation is a list ranked by similarity. Such a retrospective screening experiment can be rated by the enrichment factor, *ef* [Eq. (1)].[5,6] A method that is superior to a random selection of compounds returns an $ef > 1$.

$$ef = \left(\frac{S_{\text{act}}}{S_{\text{all}}}\right) / \left(\frac{P_{\text{act}}}{P_{\text{all}}}\right) \quad (1)$$

P_{all} is the total number of compounds in the database ("pool"), and S_{all} is the number of molecules in the subset. P_{act} is the number of "active" molecules in the pool, and S_{act} is the number of actives found in the subset.In this study, we examined the influence of seven similarity measures on the enrichment of actives using a pharmacophore-based correlation vector descriptor and 12 different datasets. In particular, we evaluated to what extent different similarity measures complement each other in terms of the retrieved active compounds. The knowledge gained will provide a basis for prospective similarity searching studies.

All molecules were extracted from the COBRA database (version 2.1), which is a collection of reference molecules for ligand-based library design compiled from recent scientific literature.[7] Twelve subsets were compiled from the 4705 COBRA compounds, containing ligands that bind to angiotensin converting enzyme (ACE, 44 compounds), cyclooxygenase 2 (COX2, 93 compounds), corticotropin releasing factor (CRF) antagonists (63 compounds), dipeptidyl-peptidase (DPP) IV (25 compounds), G-protein coupled receptors (GPCR, 1642 compounds), human immunodeficiency virus protease (HIVP, 58 compounds), nuclear receptors (NUC, 211 compounds), matrix

[a] *U. Fechner, Prof. Dr. G. Schneider*
Johann Wolfgang Goethe-Universität
Institut für Organische Chemie und Chemische Biologie
Marie-Curie-Straße 11, 60439 Frankfurt (Germany)
Fax: (+49) 69-79829826
E-mail: gisbert.schneider@modlab.de

DOI: 10.1002/cbic.200300812

metalloproteinase (MMP, 77 compounds), neurokinin (NK) receptors (188 compounds), peroxisome proliferator-activated receptor (PPAR, 35 compounds), beta-amyloid converting enzyme (BACE, 44 compounds), and thrombin (THR, 188 compounds). All molecules of one subset were considered as "active" at a time, and the respective remainder of the COBRA database as "inactive". The choice of these subsets was based on different levels of specificity. This means that we included sets of ligands binding to individual receptor subtypes (e.g., BACE, THR) as well as very loosely defined classes of bioactive compounds (e.g., GPCR, NUC). This concept reflects the idea of current chemogenomics approaches that try to design focused libraries at such different levels of specificity.[8] The group of nuclear receptors, for example, consists of more than 60 proteins like PPAR, CAR, LXR, and FXR, and comprises a variety of biological functions.[9] It can be useful to design compound libraries even at this coarse level in order to de-orphanize receptors of the same family or identify multiple compound activities within the same receptor class.[10]

In the present study, compounds were encoded by the CATS descriptor, which belongs to the category of atom-pair descriptors,[11] and encodes topological pharmacophore information.[12]

A lot of different similarity metrics exist. In this study, seven such metrics were compared (Table 1). We selected six established metrics that are frequently employed for chemical similarity searching and added the spherical distance to this list because it has not been used for this purpose before. As the CATS descriptor is composed of non-binary values, the formulae for continuous variables were employed for all similarity metrics.[13]

Table 1. *Similarity measures for continuous variables.* A *and* B *are objects (here: molecules),* i *and* j *are attributes of these objects,* n *is the total number of attributes of an object,* x_{jA} *the value of the jth attribute of object* A, $S_{A,B}$ *denotes the similarity between objects* A *and* B, *and* $D_{A,B}$ *the distance between objects* A *and* B.

Name	Equation	Range
Manhattan distance	$D_{A,B}=\sum_{j=1}^{j=n} \lvert x_{jA}-x_{jB}\rvert$	0 to ∞
Euclidian distance	$D_{A,B}=\sqrt{\sum_{j=1}^{j=n} (x_{jA}-x_{jB})^2}$	0 to ∞
Tanimoto coefficient	$S_{A,B}=\frac{\sum_{j=1}^{j=n} x_{jA}x_{jB}}{\sum_{j=1}^{j=n}(x_{jA})^2+\sum_{j=1}^{j=n}(x_{jB})^2-\sum_{j=1}^{j=n}x_{jA}x_{jB}}$	−0.33 to +1
Soergel distance	$D_{A,B}=\frac{\sum_{j=1}^{j=n}\lvert x_{jA}-x_{jB}\rvert}{\sum_{j=1}^{j=n}\max(x_{jA},x_{jB})}$	0 to 1
Dice coefficient	$S_{A,B}=\frac{2\sum_{j=1}^{j=n}x_{jA}x_{jB}}{\sum_{j=1}^{j=n}(x_{jA})^2+\sum_{j=1}^{j=n}(x_{jB})^2}$	−1 to +1
Cosine coefficient	$S_{A,B}=\frac{\sum_{j=1}^{j=n}x_{jA}x_{jB}}{\sqrt{\sum_{j=1}^{j=n}(x_{jA})^2\sum_{j=1}^{j=n}(x_{jB})^2}}$	−1 to +1
Spherical distance	$D_{A,B}=\alpha\cos(A_N{*}B_N)^{[a,b]}$	0 to π

[a] $A_N=\frac{A}{\lvert A\rvert}=\sum_{j=1}^{j=n}\left[\frac{x_{jA}}{\sqrt{\sum_{i=1}^{i=n}x_{iA}^2}}\right]$. [b] $B_N=\frac{B}{\lvert B\rvert}=\sum_{j=1}^{j=n}\left[\frac{x_{jB}}{\sqrt{\sum_{i=1}^{i=n}x_{iB}^2}}\right]$.

We calculated the enrichment factors for the 12 datasets from the first five percent (235 compounds) of the screened database. Apart from in the GPCR dataset, we were able to considerably increase the percentage of active compounds, obtaining an average $ef \geqslant 4$. The greatest enrichment factor ($ef = 12$) was yielded for the ACE dataset with the Soergel distance. This might be a consequence of the close structural similarity of the ACE ligands that were used as reference. In contrast, the GPCR set represented a very loosely defined compilation of molecules containing modulators of all classes of GPCRs.

Enrichment factors for the same dataset but different similarity metrics varied only slightly. The deviations ranged from zero (DPP-IV, GPCR, MMP, NK, and BACE) to a maximum of two units (ACE and HIVP). For almost all datasets, the Manhattan and the Soergel distances yielded the overall highest enrichment factors. It must be stressed that the homogeneity of the enrichment factors of different similarity measures does not reveal information about the homogeneity of the retrieved active molecules. The enrichment factor discriminates only between "active" and "inactive". Thus it is not advisable to compare the performance of different similarity searching methods exclusively by means of the enrichment factor. The enrichment factor represents a measure for quantification of a similarity search, but it does not comprise qualitative considerations. Therefore, we then analyzed which active molecules were retrieved by each similarity metric.

Retrospective screening with the seven different similarity metrics yielded seven similarity-ranked lists. The active compounds ("hits") that were found within the first 5% of each list were extracted. Then we stepwise united the hits found in the individual lists. This procedure led to the retrieval of significantly more hits than found by any single similarity metric. Figure 1 illustrates this gradual rise for the NUC dataset. Sixteen of these actives were retrieved by each similarity metric, examples are darglitazone **1**, a PPAR-γ ligand,[14] and com-

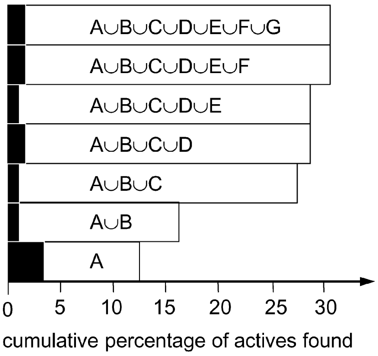
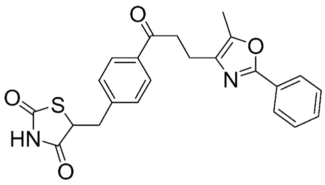
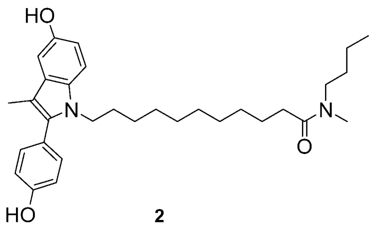


Figure 1. *The "cumulative percentage" of active compounds found among the top-ranking 5% of the similarity-ranked list that results from a retrospective screening (white bars). The NUC subset of the COBRA database was selected as an example. A) Manhattan distance, B) Euclidian distance, C) Tanimoto coefficient, D) Soergel distance, E) Dice coefficient, F) Cosine coefficient, G) Spherical distance. Black bars show the percentage of actives that were retrieved by the respective similarity metric and no more than one additional similarity metric.*

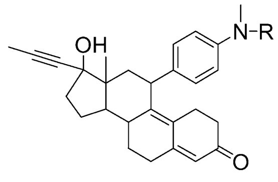
pound **2**, an estrogen receptor ligand.[15] In contrast, structures **3** (Mifepristone; an antiprogestine and glucocorticoid receptor antagonist; R: bile acid conjugate),[16] **4** (LG-100364, a "rexinoid" binding to the PPAR-γ/RXR heterodimer),[17] and **5** (Adapalene, a selective retinoic acid receptor β antagonist)[18] represent hits that were mutually retrieved only by the cosine coefficient and the spherical distance. None of the other similarity metrics claimed these known nuclear receptor ligands among the top five percent of the ranked lists. This demonstrates that it might be wise to combine individual results for a focused library rather than considering only those compounds that are jointly retrieved by each metric. It should be stressed that such a retrospective screening with a heterogeneous set of query structures, as in our NUC example, will result in a focused library reflecting the same level of specificity. In this study it was demonstrated that this approach is applicable to the design of libraries that are focused on a receptor class and not a single receptor subtype.



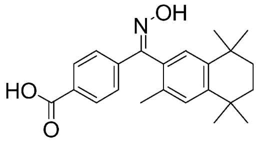
1



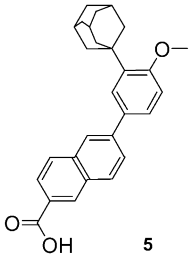
2



3



4



5

"Cumulative percentages" for all seven similarity metrics allowed for the retrieval of up to 74% of the actives in the first five percent of the database in the case of ACE. The increase of the cumulative percentages for all seven metrics compared to the employment of only the Manhattan distance ranged from additional 5% (COX2) to 28% (NUC and MMP) with an average of 19% over all 12 datasets.

Each similarity metric alone proved to be able to retrieve and increase the percentage of active compounds in a focused library. Nevertheless, their respective definition of "distance" in chemical space (in this study, a topological pharmacophore space) has a strong impact on the structural diversity among the highest-ranking active compounds. This is illustrated by black bars in Figure 1. These black bars indicate the percentage of active compounds that were retrieved by both the respective similarity metric and exactly one other metric. It should be stressed that the example shown in Figure 1 does not justify general conclusions to be drawn concerning the usefulness of a similarity metric, as their pair-wise performance differed for the various sets of ligands investigated here. We conclude that different similarity metrics complement each other. Therefore, it might be advantageous to employ several molecular descriptors and similarity metrics in parallel and benefit from a unification of the various definitions of "chemical similarity". This idea supplements earlier findings by Bradshaw and co-workers who demonstrated that combining different similarity metrics can lead to improved enrichment of actives by applying "fusion rules" to the individual ranked lists.[19] In the present study we demonstrate that unification of the sets of highest-ranking compounds that were retrieved with the different similarity metrics can serve the same purpose.

Acknowledgements

The authors are grateful to Norbert Dichter for setting up the LSF Linux cluster and to Petra Schneider for compiling the COBRA sets. This work was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt.

Keywords: computer chemistry · drug design · libraries · pharmacophore · virtual screening

[1] J. M. Barnard, G. M. Downs, P. Willett in *Virtual Screening of Bioactive Molecules* (Eds.: H. J. Böhm, G. Schneider), Wiley-VCH, Weinheim, **2000**, pp. 59–80.
[2] G. Schneider, M. Nettekoven, *J. Comb. Chem.* **2003**, *5*, 233–237.
[3] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
[4] M. Stahl, M. Rarey, G. Klebe in *Bioinformatics: From Genomes to Drugs, Vol. 2*, (Ed.: T. Lengauer), Wiley-VCH, Weinheim, **2001**, pp. 137–170.
[5] H. Xu, D. K. Agrafiotis, *Curr. Top. Med. Chem.* **2002**, *2*, 1305–1320.
[6] U. Fechner, L. Franke, S. Renner, P. Schneider, G. Schneider, *J. Comput. Aided Mol. Des.* **2003**, *17*, 687–698.
[7] P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713–718.
[8] A. W. E. Chan, J. P. Overington, *Annu. Rep. Med. Chem.* **2003**, *38*, 285–294.
[9] T. M. Willson, P. J. Brown, D. D. Sternbach, B. R. Henke, *J. Med. Chem.* **2000**, *43*, 527–550.
[10] K. H. Bleicher, H. J. Böhm, K. Müller, A. I. Alanine, *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.
[11] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem.* **1999**, *111*, 3068–3070; *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896.
[12] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
[13] P. Willett, J. M. Barnard, G. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
[14] L. Pickavance, P. S. Widdowson, P. King, S. Ishii, H. Tanaka, G. Williams, *Br. J. Pharmacol.* **1998**, *125*, 767–770.
[15] C. P. Miller, B. S. Komm, *Annu. Rep. Med. Chem.* **2001**, *36*, 149–158.
[16] R. Rupprecht, J. M. Reul, B. van Steensel, D. Spengler, M. Soder, B. Berning, F. Holsboer, K. Damm, *Eur. J. Pharmacol.* **1993**, *247*, 145–154.
[17] J. Cobb, I. Dukes, *Annu. Rep. Med. Chem.* **1998**, *33*, 213–222.
[18] A. M. Nadzan, *Annu. Rep. Med. Chem.* **1995**, *30*, 119–128.
[19] C. M. R. Ginn, P. Willett, J. Bradshaw, *Perspec. Drug Discovery Des.* **2000**, *20*, 1–16.

Received: November 5, 2003 [Z 812]