A Distance Function for Retrieval of Active Molecules from Complex Chemical Space Representations

Jeffrey W. Godden and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2,
D-53113 Bonn, Germany

Received November 22, 2005

The concept of chemical space is of fundamental importance for chemoinformatics research. It is generally thought that high-dimensional space representations are too complex for the successful application of many compound classification or virtual screening methods. Here, we show that a simple "activity-centered" distance function is capable of accurately detecting molecular similarity relationships in "raw" chemical spaces of high dimensionality.

INTRODUCTION

Computational approaches such as diversity analysis, library design, compound classification, or virtual screening depend on chemical reference spaces that are formed by different molecular descriptors.[1,2] Typically, each of *n* chosen descriptors adds one dimension (or axis) to an *n*-dimensional chemical space. The high dimensionality of chemical reference spaces and descriptor correlation effects have hampered many computational applications and complicated the analysis of their results.[1,2] Consequently, much emphasis has been put on the development of methods to reduce the dimensionality and complexity of chemical spaces or remove distortions arising from descriptor correlation effects.[1−3] Similarly, techniques are available to generate low-dimensional and orthogonal spaces a priori[3,4] or, alternatively, simplify descriptor representations for space design.[5] Preferential operation in low-dimensional and orthogonal chemical space representations has become a paradigm in chemoinformatics research,[2−4] and methods that depart from this theme are rare.[5−7]

Previously, we have developed partitioning methods that utilize binary-transformed property descriptors[5] for the generation of simplified high-dimensional reference spaces.[5,6] Here, we address the difficulties associated with the use of high-dimensional chemical spaces by developing a simple distance function (distance in activity-centered chemical space; DACCS) for compound classification and ligand-based virtual screening. The DACCS approach directly operates in high-dimensional descriptor spaces and is designed to be a compound ranking method. Thus, it produces a distance-based ranking of database compounds as a measure of similarity to known active molecules. The design and initial evaluation of DACCS is reported herein.

METHODS AND CALCULATIONS

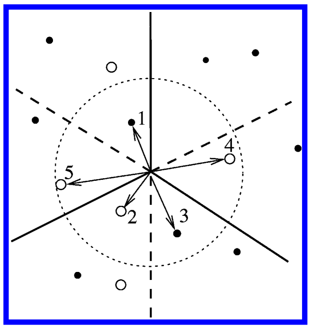

Figure 1. DACCS approach. High-dimensional chemical space is centered on a subspace where known compounds with similar activity concentrate. The calculation of Euclidian-like distances from the "activity center" for all database compounds produces a distance-based ranking that is used as a measure of molecular similarity. Open dots represent active compounds, and black dots represent other database compounds. The dashed circle indicates the "activity radius", as discussed in the text. Arrows represent distances, and 1−5 are the top five compounds in a distance-based ranking for an activity class.

A schematic representation summarizing the ideas underlying the DACCS approach is shown in Figure 1. Through a scaling procedure, the method centers chemical spaces produced by an arbitrary number of descriptors on a subspace that is populated by a set of active compounds and then superimposes an orthogonal grid as an approximation onto this subspace. This is done by the calculation of Euclidian-like distances from the center of the subspace to all database compounds in chemical space, thereby producing a distance-based ranking corresponding to decreasing similarity. Unique features of DACCS include that it operates directly in chemical space and it takes all descriptor contributions with detectable information content into account. Because of its conceptual simplicity, DACCS is easily applicable to high-dimensional space representations.

Given a set of known active compounds (templates or "baits"), the distance in scaled chemical space (d_{DACCS}) from the center of the "active subspace" is calculated for database compounds as follows:

$$d_{\mathrm{DACCS}} = \sqrt{\sum_{i}^{d} \left(\frac{x_i - \overline{\mathrm{act}_i}}{\mathrm{stdev}_i} \right)^2}$$

Here, x_i is the value of one of *d* descriptors for a compound

* Corresponding author tel.: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci050510i CCC: $33.50 © 2006 American Chemical Society
Published on Web 03/28/2006

Table 1. Virtual Screening Trials Using DACCS[a]

			average ADC found				recovery rate (%)				hit rate (%)			
class	bait	ADC	S10	S25	S50	S100	S10	S25	S50	S100	S10	S25	S50	S100
TKE	10	10	6.0	6.6	7.1	7.8	60.4	66.4	71.2	77.6	60.4	26.6	14.2	7.8
COX	10	7	6.0	6.2	6.2	6.2	86.3	88.0	88.0	88.0	60.4	24.6	12.3	6.2
CAE	10	12	4.1	4.6	5.3	6.0	34.0	38.3	44.0	49.7	42.5	19.2	11.0	6.0
H3E	10	11	5.5	6.7	7.5	7.9	50.2	61.1	68.4	72.0	55.2	26.9	15.0	7.9
HIV	10	8	1.5	2.0	2.2	2.4	18.5	24.5	27.5	30.5	14.8	7.8	4.4	2.4

[a] Compound activity classes are abbreviated as follows: TKE, tyrosine kinase inhibitors (20 compounds); COX, cyclooxygenase-2 inhibitors (17); CAE, carbonic anhydrase II inhibitors (22); H3E, H3 antagonists (21); and HIV, HIV protease inhibitors (18). "Bait" reports the number of template compounds used for each calculation, and "ADC" stands for active database compounds (potential hits). Under "average ADC found", the columns "S10", S25", "S50", and "S100" report the number of hits identified within selection sets of either 10, 25, 50, or 100 database compounds most proximal (similar) to the baits. The "recovery rate" is defined as the number of ADC (hits) found divided by the total number of ADC in the database, and "hit rate" is the number of ADC divided by the total number of selected database compounds. For each activity class, average values are reported for 25 independent virtual screening trials.

Table 2. Similarity Search Calculations Using a Structural Key Fingerprint[a]

			average ADC found				recovery rate (%)				hit rate (%)			
class	bait	ADC	S10	S25	S50	S100	S10	S25	S50	S100	S10	S25	S50	S100
TKE	1	19	1.7	2.5	3.0	3.8	8.9	13.4	15.8	20.0	17.0	10.2	6.0	3.8
COX	1	16	1.8	2.2	2.4	2.7	11.0	13.6	15.1	16.9	17.6	8.7	4.8	2.7
CAE	1	21	2.5	3.8	4.7	5.3	12.1	18.2	22.3	25.1	25.5	15.3	9.4	5.3
H3E	1	20	5.0	6.4	6.6	7.2	25.0	31.9	33.1	36.0	50.0	25.5	13.2	7.2
HIV	1	17	5.2	6.0	6.4	6.7	30.7	35.3	37.9	39.5	52.2	24.0	12.9	6.7

[a] For each activity class, one compound at a time was used as the bait, and the remaining active compounds were added to the database as potential hits (ADC). In analogy to the DACCS calculations in Table 1, the number of hits within the top-ranked 10, 25, 50, and 100 database compounds is reported. For each activity class, the data shown represent the average over all search calculations.

x, $\overline{\mathrm{act}}_i$ is the mean value of descriptor i for a set of active template compounds, and stdev_i is the standard deviation of the descriptor values for the templates. If this standard deviation is zero, then the standard deviation of the entire compound population is used instead, which effectively gives such descriptors a unit weighting relative to the database compounds. If both values are zero, the descriptor is omitted from the calculations. This applies to a descriptor having the same value for all active and database compounds and thus no information content.

In addition to these calculation parameters, we define the "activity radius" (Act_Rad) of a template set as the mean distance of all active compounds from the center of the set in descriptor space. Because of the use of the standard deviation for distance scaling, Act_Rad is equivalent to the square root of the number of descriptors used. However, because of the possibility that standard deviations of the compound population might also be used, the equivalence is only approximate.

In ligand-based virtual screening, one attempts to use known active compounds as "baits" in order to identify novel hits in compound databases.[2] Therefore, we have tested the DACCS function on five previously assembled compound activity classes (different inhibitors and antagonists)[8] that are reported in Table 1 and contain between 17 and 22 compounds. For each class, 25 sets of 10 bait compounds each were randomly selected, and the remaining compounds were "hidden" as potential hits in a background database of ∼1.34 million molecules collected from medicinal chemistry sources. With the exception of known active molecules, all database compounds were considered inactive and potential false positives in DACCS calculations. Thus, only very few potential hits were available within a large number of background molecules, providing challenging conditions for virtual screening.

As reference calculations, we have carried out similarity searches in the background database using a fingerprint consisting of the publicly available set of 166 MACCS structural keys.[9] In these calculations, each active compound was used once as a template, and the remaining active molecules were added to the database as potential hits. Database compounds were ranked by Tanimoto coefficient (Tc)[10] relative to each bait molecule, and for each activity class, the results were averaged over all search calculations.

As descriptors, we used a set of 123 previously described 1D, 2D, and implicit 3D descriptors[5] available in the Molecular Operating Environment.[11] Consistent with the basic ideas of DACCS, we intended to use as many descriptors as possible. Our only requirement was that their values should not depend on hypothetical compound conformations. While many descriptors are expected to display a degree of correlation, this is thought to contribute only a small amount of relative descriptor weighting as compared to the primary weighting resulting from relative variance.

We also tested the ability of DACCS to detect active molecules within the activity radius of each compound class. Therefore, activity classes were added to the Molecular Drug Data Report (MDDR)[12] containing ∼160 000 biologically active compounds, and the number of MDDR compounds falling into the activity radius of each class was determined and their activities were analyzed.

RESULTS AND DISCUSSION

The results of our simulated virtual screening trials are summarized in Table 1. In these calculations, the average dimensionality of the descriptor spaces was 122. For each

Table 3. Descriptor Variances[a]

class	D_p ≥ 0.05	D_p < 0.05	diff. (%)
TKE	16	107	87.0
COX	15	108	87.8
CAE	18	105	85.4
H3E	11	112	91.1
HIV	39	84	68.3

[a] *P* values were calculated for a comparison of descriptor variances within each activity class and the background database. "D_p ≥ 0.05" reports the number of descriptors with *P* values equal to or greater than 5%, and "D_p < 0.05" reports the number of descriptors with *P* values smaller than 5% that are considered to reflect significant statistical differences; "diff." reports the percentage of descriptors displaying such differences.

individual calculation, we selected differently sized sets of database compounds on the basis of closest distances in descriptor space to the center of each activity class and determined recovery and hit rates. Table 1 shows that DACCS produced encouraging results in these trials. For example, for small selection sets consisting of 25 compounds, average recovery rates between ∼25% (HIV) and 88% (COX) and hit rates between ∼8% (HIV) and 27% (H3E) were achieved. It should be noted that, for the selection of 25 database compounds, the highest possible hit rate in any of these calculations was only 48% (12 potential hits). For the selection of only 10 database compounds, all classes except HIV produced very significant hit rates between ∼43% and 60%.

To compare the DACCS results with a standard similarity-based method, reference calculations were carried out using a structural key fingerprint. DACCS and fingerprint search results can be well compared because both methods produce a similarity ranking (DACCS by distance, fingerprints by Tc values) of database compounds relative to baits. The results are reported in Table 2. In these search calculations, for each activity class, there were about twice as many potential hits available in the background database as for DACCS analysis. Nevertheless, DACCS produced consistently and significantly higher hit and recovery rates for four of five activity classes, except for HIV, where similarity searching produced better results. For compound selection sets consisting of 10 or 25 compounds averaged over all of the activity classes, DACCS produced hit and recovery rates of 33.8% and 52.7%, respectively, whereas fingerprint searching produced average hit and recovery rates of 24.6% and 20.0%, respectively.

Furthermore, activity classes TKE, COX, and H3E were previously included in benchmark sets for two other virtual screening methods that operate in simplified descriptor spaces, dynamic mapping of consensus positions (DMC)[6] and recursive median partitioning (RMP).[13] On these activity classes, DACCS calculations achieved consistently higher hit and recovery rates than RMP, which produced maximum hit and recovery rates of 13% and 40%, respectively.[13] Recovery rates were also consistently better for DACCS than for DMC, where a maximum recovery rate of 29% was observed for H3E.[6] By contrast, DACCS and DMC achieved hit rates comparable in magnitude (with DACCS producing better results on TKE and COX and DMC producing better results on H3E).

Table 3 reports the results of a statistical variance test comparing descriptor variances within activity classes and the background database. For all activity classes, the majority of descriptors have variances that significantly differ from the background database. The percentage of descriptors with such differences in variance is smallest for class HIV where DACCS produced the lowest hit and recovery rates.

We have also carried out control calculations with DACCS where only the descriptor standard deviations in the background database were used for scaling. This modification removes the influence of activity-class-specific descriptor variances and distorts the activity-dependent subspaces. The results are reported in Table 4. While DACCS calculations detected on average close to five active molecules among the top 10 ranked database compounds (Table 1), on average, fewer than two active compounds were identified in these

Table 4. DACCS Calculations Using Only Database Standard Deviations as Weights[a]

			average ADC found				recovery rate (%)				hit rate (%)			
class	bait	ADC	S10	S25	S50	S100	S10	S25	S50	S100	S10	S25	S50	S100
TKE	10	10	1.2	1.2	1.3	1.4	12.0	12.4	13.1	13.8	12.1	4.8	2.6	1.4
COX	10	7	2.2	2.6	3.4	3.8	31.4	37.2	48.6	54.3	22.2	10.4	6.8	3.8
CAE	10	12	0.8	1.2	1.2	1.3	6.7	10.0	10.0	10.8	8.4	4.8	2.4	1.3
H3E	10	11	4.0	4.4	5.4	5.6	36.3	40.0	49.1	50.7	40.3	17.6	10.8	5.6
HIV	10	8	1.4	1.6	1.6	1.7	17.5	20.0	20.0	21.2	14.3	6.4	3.2	1.7

[a] Results are reported and abbreviations used according to Table 1.

Table 5. Screening by Activity Radius

class	D_bait	D_db	Act_Rad	DB_Rad	DB_Rad*	hit rate (%)	hit rate* (%)
TKE	105	17	10.00	4	7	100	100
COX	108	14	10.08	14	19	93	95
CAE	110	12	10.25	39	39	54	54
H3E	109	13	10.19	2	2	100	100
HIV	110	12	10.19	84	87	55	54

[a] "D_bait" reports the number of descriptors for which the standard deviation for the bait set was not equal to zero, and "D_db" reports the number of descriptors for which the database standard deviation was used instead. "Act_Rad" is the activity radius. "DB_Rad" reports the number of database compounds falling within the activity radius, and "hit rate" reports the percentage of those compounds having the same activity designation as the baits. "DB_Rad*" reports the number of database compounds within the activity radius after omitting activity-class-invariant descriptors (i.e. D_db) from the calculations, and "hit rate*" reports the corresponding hit rates.

control calculations, leading to significant reductions in hit and recovery rates. These findings confirm the relevance of the activity-centered scaling procedure of DACCS.

In addition to simulated virtual screening trials, DACCS was applied to screen the MDDR in order to detect active molecules falling into the activity radius of each class. Only two of our bait molecules were present in this database and were removed from it. For each class, we calculated the number of database compounds falling into its activity radius and determined their activities. The results reported in Table 5 are also of considerable interest. When 122 descriptors were utilized, only small numbers of database compounds (with a maximum of 84 for HIV) fell into the activity radii of the activity classes. These findings further supported a high selectivity of DACCS calculations. Importantly, dependent on the activity class, 50−100% of the database compounds falling into an activity radius had the same activity as the templates (Table 5). These results demonstrated the ability of the DACCS function to accurately detect molecular similarity relationships in “unrefined” chemical reference spaces of high dimensionality. In addition, compound distributions within the activity radii have been calculated after omitting activity-class-invariant descriptors. The results are also reported in Table 5. Omitting invariant descriptors does not change an activity radius because those descriptors map to the centroid and, thus, do not contribute to distances. However, omitting these descriptors has a potential effect on the ensemble of compounds falling within the unchanged radius of activity because the dimensionality of the reference space changes. The data in Table 5 show that the compound ensemble only changes for two of five activity classes and that these changes are subtle. Thus, activity-class-invariant descriptors do not strongly influence the results of DACCS calculations.

When designing a high-dimensional compound selection technique, one must consider the “curse of dimensionality”. As the dimensionality of the space representation increases, a growing proportion of the chemical reference space falls outside of the resulting hypersphere and into the corners of the enclosing hypercube. However, for mining compound databases of large size, such effects can be helpful because they lead to a reduction in the number of candidate compounds that need to be considered.

The DACCS approach is characterized by its methodological simplicity and significant predictive ability, as suggested by the results obtained in our pilot calculations. Our findings show that it is possible to establish meaningful distance relationships between compounds in complex space representations. When the DACCS function has been applied, proximity in complex chemical space has successfully been used as a measure of biological activity, in a similar way to what has been accomplished, for example, by nearest-neighbor searching in low-dimensional reference spaces.[14] Our results indicate that methods of high computational complexity are not required to successfully operate in high-dimensional spaces and detect molecular similarity relationships.

CONCLUSIONS

We have reported a molecular similarity method designed to identify active compounds in high-dimensional chemical space representations. The DACCS approach is characterized by its simplicity. As a distance function, DACCS does not require the careful preselection of molecular descriptors for the generation of reference spaces, which greatly simplifies space design. A key finding has been that DACCS distance calculations successfully rank compounds according to biological activity in high-dimensional descriptor spaces. In addition, we have been able to show that DACCS calculations produce significant hit and recovery rates in virtual screening trials.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, R. F. Combinatorial Informatics in the Post-Genomics Era. *Nat. Rev. Drug Discovery* **2002**, *1*, 337−346.

(2) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(3) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.

(4) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 28−35.

(5) Godden J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 885−893.

(6) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular Similarity Analysis and Virtual Screening in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21−29.

(7) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(8) Xue, L.; Bajorath, J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 757−764.

(9) *MACCS Structural Keys*; MDL Information Systems Inc.: San Leandro, CA.

(10) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(11) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada.

(12) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA.

(13) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive Median Partitioning for Virtual Screening of Large Databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.

(14) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-Directed Nearest Neighbor Searching. *J. Med. Chem*. **2005**, *48*, 240−248.

CI050510I