modlab

Distance Phenomena in Chemical Spaces: Consequences for Similarity Approaches

Matthias Rupp[1], Petra Schneider[2], Gisbert Schneider[1]
matthias.rupp@bio.uni-frankfurt.de

[1] Beilstein-Endowed Chair for Cheminformatics, Johann Wolfgang Goethe-University, Siesmayerstr. 70, 60323 Frankfurt, Germany
[2] Schneider Consulting GbR, George-C.-Marshall Ring 33, 61440 Oberursel, Germany

FIRST Frankfurt International Research Graduate School for Translational Biomedicine

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

1 Introduction

- (Dis)similarity measures are important for applications like compound ranking, clustering, property prediction, virtual screening, and diversity analysis.
- Molecular descriptors constitute (high-dimensional) vector spaces (Table 1).
- Metrics, inner products and similarity coefficients measure (dis)similarity.

- Which measure is best for a given task?
- Is high dimensionality problematic?

- Norms measure length (size), e.g., L^p-norms $||x||_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ (Figs. 1,2).
- Metrics measure distance, e.g., Minkowski metrics $||x,y||_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{1/p}$, with special cases Manhattan (p=1), Euclidean (p=2) and max (p=∞) distance.
- Inner products measure length and angle, e.g., $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$.
- Similarity coefficients are often normalized inner products, e.g., Tanimoto coefficient $\frac{\langle x,y\rangle}{\langle x,x\rangle + \langle y,y\rangle - \langle x,y\rangle} \in [-\frac{1}{3}, 1]$ and Carbó index $\frac{\langle x,y\rangle}{||x||_2\,||y||_2} \in [-1, 1]$.

- Norms, metrics and inner products [1] generalize geometric concepts like length, distance, angle and orthogonality to higher dimensions.
- Measurement is across volume, which grows exponentially with dimension.
- This causes phenomena in higher dimensions which influence computations. Implications were investigated on artificial and biochemical data.
- The COBRA [2] drug dataset, version 8.6 (10,886 annotated compounds) was used with the CATS2D [3] and all MOE [4] 2D descriptors. For classification, a subset of 96 classes (>30 members in each) based on target and interaction type was used.

Table 1: Dimensionality d of several common descriptor spaces.

d	Descriptor
50	Mini-fingerprints [5]
72	VolSurf [6]
120	Ghose-Crippen [7]
150	CATS2D [3]
184	MOE 2D [4]

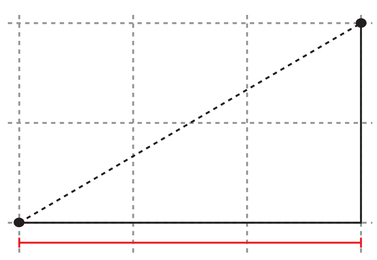


Figure 1: Minkowski distances $||\cdot||_1$ (solid), $||\cdot||_2$ (dashed) and $||\cdot||_\infty$ (red).

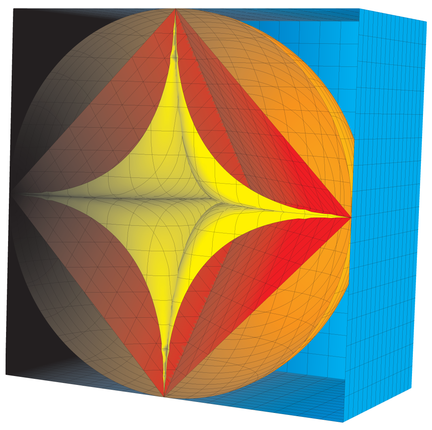


Figure 2: Unit sphere using (from inside to outside) $||\cdot||_{0.5}$, $||\cdot||_1$, $||\cdot||_2$, and $||\cdot||_\infty$.

2 Distance Phenomena

Empty Space Phenomenon

- Dividing each dimension into two parts yields 2^d compartments.
- n samples can span max(n-1,d) dimensions, but they can cover only $\log_2 n$ dimensions, in the sense that all compartments contain at least one sample.
- Most of chemical descriptor space is empty.
- For independent uniform samples, the probability of at least one shared compartment is $1 - \binom{m}{n}\frac{n!}{m^n}$ with $m=2^d$. For COBRA, this is $<10^{-34}$, but >1000 compartments are actually shared.
- This indicates a lower intrinsic dimensionality of the data (Fig. 3).

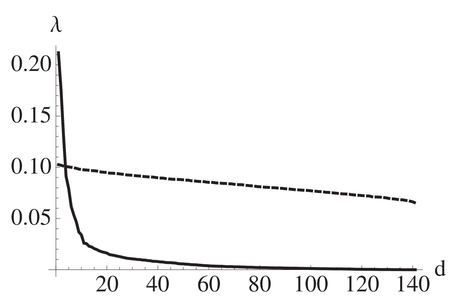


Figure 3: Principle component analysis eigenvalues, COBRA data using CATS2D (solid) and random samples (dashed).

Sphere Phenomena

- The d-dimensional Euclidean unit sphere (Fig. 2) has volume $V_{S_{d,r}} = (\pi^{d/2} r^d)/\Gamma(1+\frac{d}{2})$, which peaks at d=5 and vanishes for $d\to\infty$.
- For $d\to\infty$, samples drawn from a sphere enclosing another one lie outside of the smaller one. For multivariate Gaussians, most probability is in the tails [8].

Distance Concentration

- Sample norms concentrate for high-dimensional data [9–11] (Fig. 4).
- As a consequence, all distances are similar, samples lie on a hypersphere, and each sample is nearest neighbor of all other samples.
- Concentration can be measured using the variation coefficient $\frac{\sigma}{\mu} = \frac{\sqrt{\mathrm{var}(||\cdot,\cdot||)}}{\mathrm{E}(||\cdot,\cdot||)}$.

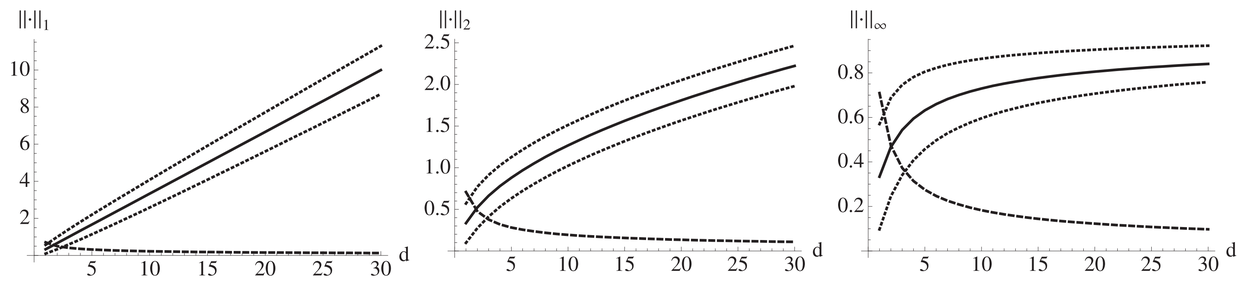
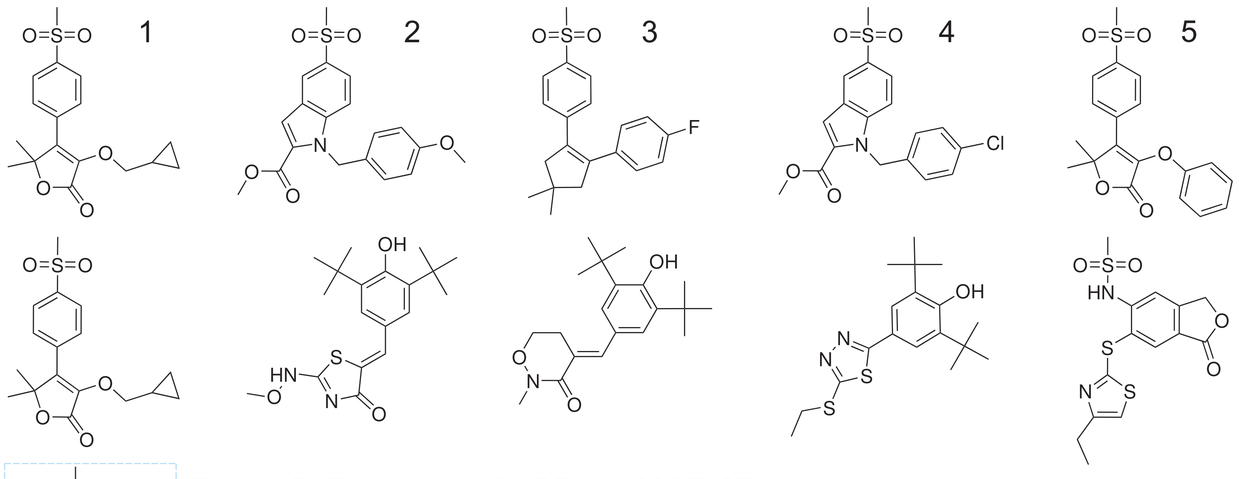


Figure 4: Minkowski metrics $||\cdot||_1$ (left), $||\cdot||_2$ (middle), and $||\cdot||_\infty$ (right) between 10^5 random samples. Mean (solid) ± standard deviation (dotted), and variation coefficient (dashed).

3 Consequences

- The choice of (dis)similarity measure influences compound ranking (Fig. 5).



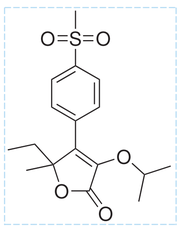


Figure 5: Ranking of COBRA & MOE 2D versus reference compound (left), using Euclidean distance (top row) and Tanimoto coefficient (bottom row).

- Due to distance concentration, inter- and intra-class distances are similar (Fig. 6), aggravating nearest neighbor clustering.

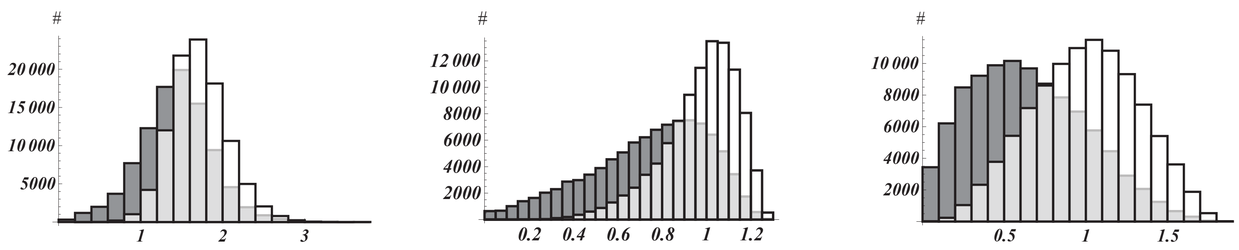


Figure 6: Intra-class (dark) and inter-class (light) distance histograms of COBRA dataset using CATS2D/Euclidean distance (left), MOE 2D/Tanimoto coefficient (middle) and MOE 2D/Carbó index (right).

4 Conclusions

- The described phenomena set in early, from 5–20 dimensions onwards. Table 1 indicates that they might be rule rather than exception.
- For compound ranking, one should use more than one (dis)similarity measure, including a Minkowski metric and a similarity coefficient (Fig. 6, Table 2, Fig. 7) to increase diversity of result lists.

cc	1	2	3	4	5	6
1	1	0.95	0.71	0.63	0.68	0.69
2		1	0.84	0.64	0.67	0.68
3			1	0.51	0.51	0.53
4				1	0.95	0.96
5					1	0.98
6						1

Table 2: Spearman rank correlation (cc) between $||.||_1$ (1), $||.||_2$ (2), $||.||_5$ (3), Pearson correlation (4), Tanimoto coefficient (5), and Carbó index (6) over COBRA / MOE 2D classes.

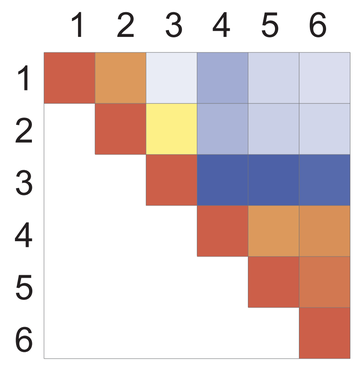


Figure 7:

- The parameter p of Minkowski metrics $||\cdot||_p$ can be determined by plotting the variation coefficient, or, for labeled data, using a maximal separation criterion.
- Empty space phenomenon, distance concentration and low intrinsic data dimensionality suggest feature selection and dimensionality reduction (Fig. 8).

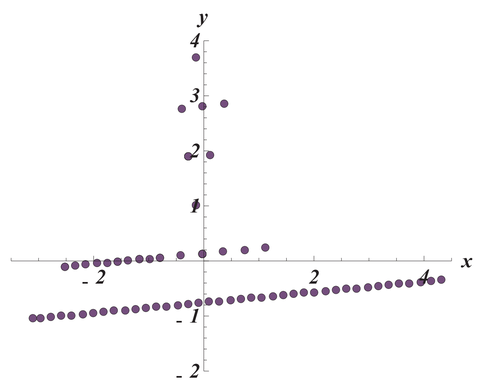


Figure 8: Dimensionality reduction example. First two principle components trained on 85 fatty acids using six descriptors. The first component (x) corresponds to carbon chain length, the second (y) to its saturation.

- More advanced concepts exist. In distance metric learning [12], the weight matrix M of the Mahalanobis distance $\sqrt{(x-y)^T M (x-y)}$ is adapted to the data to maximize class separability.

References

[1] Meyer, C.: Matrix Analysis and Applied Linear Algebra, SIAM, 2001.
[2] Schneider, P., Schneider, G.: Collection of Bioactive Reference Compounds for Focused Library Design, QSAR Comb. Sci. 22(7): 713, 2003.
[3] Schneider, G., Neidhart, W., Giller, T., Schmid, G.: "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening, Angew. Chem. Int. Ed. 38(19): 2894, 1999.
[4] Molecular Operating Environment, version 2007.09, Chemical Computing Group, www.chemcomp.com, 2008.
[5] Xue, L., Godden, J., Bajorath, J.: Evaluation of Descriptors and Mini-Fingerprints for the Identification of Molecules with Similar Activity, J. Chem. Inform. Comput. Sci. 40(5): 1227, 2000.
[6] Cruciani, G., Crivori, P., Carrupt, P.-A., Testa, B.: Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach, J. Mol. Struct. 503(1-2): 17, 2000.
[7] Viswanadhan, V., Ghose, A., Revankar, G., Robins, R.: Atomic Physicochemical Parameters for Three Dimensional Structure Directed QSAR 4: Additional Parameters, J. Chem. Inform. Comput. Sci. 29(3): 163, 1989.
[8] Scott, D.: Multivariate Density Estimation, 1992.
[9] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "Nearest Neighbor" meaningful?, ICDT 1999, 217.
[10] Hinneburg, A., Aggarwal, C., Keim, D.: What is the Nearest Neighbor in High Dimensional Spaces?, VLDB 2000, 506.
[11] François, D., Wertz, V., Verleysen, M.: The Concentration of Fractional Distances, IEEE Trans. Knowl. Data Eng., 873, 2007.
[12] Xing, E., Ng, A., Jordan, M., Russell, S.: Distance Metric Learning, with Application to Clustering with Side-Information, NIPS 2002, 505.

Support: Beilstein Institute, FIRST.

http://www.modlab.de