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 measures coefficient (dis)similarity.

Which measure is best for a given task?

- Is high dimensionality problematic?

- Norms measure length (size), e.g., $|x|_p = \left(\sum |x_i|^p\right)^{1/p}$ (Figs. 1, 2).
- Metrics measure distance, e.g., Minkowski metrics $d(x, y) = \left(\sum |x_i - y_i|^p\right)^{1/p}$ with special cases Manhattan ($p=1$), Euclidean ($p=2$) and max ($p=\infty$) distance.

- Inner products measure length and angle, e.g., $\langle x, y \rangle = \sum x_i y_i$.
- Similarity coefficients are often normalized inner products, e.g., Tanimoto coefficient $T(x, y) = \frac{\sum x_i y_i}{\sum (x_i + y_i - x_i y_i)} \in [0, 1]$ and Carlbó index $\beta(x, y) \leq |x - y|_\infty$.

- 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 grow in influence. 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 of several common descriptor spaces.

<table>
<thead>
<tr>
<th>d</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>Mini-fingerprints [5]</td>
</tr>
<tr>
<td>72</td>
<td>VolSurf [6]</td>
</tr>
<tr>
<td>120</td>
<td>Ghose-Crippen [7]</td>
</tr>
<tr>
<td>150</td>
<td>CATS2D [3]</td>
</tr>
<tr>
<td>184</td>
<td>MOE 2D [4]</td>
</tr>
</tbody>
</table>

![Figure 1: Minkowski distances $||x||_p$ (solid), $||x||_2$ (dashed) and $||x||_\infty$ (red).](image1)

![Figure 2: Unit sphere using (from inside to outside) $||x||_0.5$, $||x||_1$, $||x||_2$, and $||x||_\infty$.](image2)

3 Consequences

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

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

![Figure 5: Ranking of COBRA & MOE 2D versus reference compound (left), using Euclidean distance (top row) and Tanimoto coefficient (bottom row).](image5)

![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/Carlbó index (right).](image6)

2 Distance Phenomena

2.1 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 is empty.

- For independent uniform samples, the probability of at least one shared compartment is $1 - \left(\frac{1}{2^d}\right)^n$.

For $d \to \infty$, $d$ samples drawn from a sphere enclosing another one lie outside of the smaller one.

For multivariate Gaussians, most probability is in the tails [8].

![Figure 3: Principle component analysis eigenvalues, COBRA data using CATS2D (solid) and random samples (dashed).](image3)

2.2 Sphere Phenomenon

- The $d$-dimensional Euclidean unit sphere has volume $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}$ 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.

![Figure 4: Minkowski metrics $||x||_1$ (left), $||x||_2$ (middle), and $||x||_\infty$ (right) between $10^5$ random samples. Mean (solid) ± standard deviation (dotted), and variation coefficient (dashed).](image4)

2.3 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 $cv = \frac{\text{mean}}{\text{sd}}$.

![Figure 5: Spearman rank correlation (cc) between $||x||_1$, $||x||_2$, $||x||_\infty$, Pearson correlation (4), Tanimoto coefficient (5), and Carlbó index (6) over COBRA / MOE 2D classes.](image5)

![Figure 6: 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.](image6)

4 Conclusions

- The described phenomena set in early, from 5–20 dimensions onwards.

- 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.

- The parameter $p$ of Minkowski metrics $||x||_p$ 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).

Table 2: Spearman rank correlation (cc) between $||x||_1$, $||x||_2$, $||x||_\infty$, Pearson correlation (4), Tanimoto coefficient (5), and Carlbó index (6) over COBRA / MOE 2D classes.

| $||x||_p$ | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------|---|---|---|---|---|---|
| 1         | 1 | 0.95 | 0.71 | 0.63 | 0.68 | 0.69 |
| 2         | 0.84 | 0.64 | 0.67 | 0.68 | 1 | 0.51 | 0.53 |
| 3         | 0.95 | 0.96 | 1 | 0.98 | 1 | 1 |

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

![Figure 7: Principle component analysis eigenvalues, COBRA data using CATS2D (solid) and random samples (dashed).](image7)

References