Implementation and Evaluation of a Docking-Rescoring Method using Molecular Footprint Comparisons

Trent E. Balius, Sudipto Mukherjee, and Robert C. Rizzo

Supplemental Material

The following corollary and theorem prove the strong functional relationship (see Figure 6) between normalized Euclidean distance and standard Pearson Correlation methods used to calculate footprint similarity described in the text. Note that the relationship is strong because the mean of the footprint vectors is usually close to zero. If the mean is not close to zero (i.e. when using a threshold-based footprint) then the functional relationship will be weaker.

Corollary 1: if $\vec{u}$ and $\vec{v}$ are unit vectors, then $\|\vec{u}-\vec{v}\| = \sqrt{2(1-\cos(\theta))}$, where θ is the angle between $\vec{u}$ and $\vec{v}$.

Proof:
Let $\vec{u}$ and $\vec{v}$ be unit vectors. Then,

$$\cos(\theta) = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\,\|\vec{v}\|} = \vec{u}\cdot\vec{v} \qquad (\because \quad \|\vec{u}\| = 1)$$

$$\|\vec{u}-\vec{v}\| = \sqrt{\sum (u_i - v_i)^2} = \sqrt{\sum {u_i}^2 + \sum {v_i}^2 - 2\sum u_i v_i}$$

$$= \sqrt{1+1-2\sum u_i v_i} \qquad (\because \quad \|\vec{u}\|^2 = \sum {u_i}^2 = 1)$$

$$= \sqrt{2(1-\vec{u}\cdot\vec{v})}$$

Therefore, $\|\vec{u}-\vec{v}\| = \sqrt{2(1-\cos(\theta))}$

Theorem: $d_{norm} \approx \sqrt{2(1-r)}$ for two vectors *x* and *y*, whose means are close to zero, where d_{norm} is the normalized Euclidean distance and r is the Pearson Correlation Coefficient (both calculated between the two vectors).

Proof:

$$r = \frac{\mathrm{cov}(\vec{x},\vec{y})}{\sqrt{\mathrm{var}(\vec{x})}\sqrt{\mathrm{var}(\vec{y})}}$$

The correlation coefficient can also be thought of as the cosine of the angle formed between the mean-modulated vectors x^{μ} and y^{μ} where $\vec{x}^{\,\mu} = [x_i - \mu_x]$ and $\vec{y}^{\,\mu} = [y_i - \mu_y]$, and μ represents the mean of each vector.

$$r = \frac{\vec{x}^{\,\mu}\cdot\vec{y}^{\,\mu}}{\|\vec{x}^{\,\mu}\|\,\|\vec{y}^{\,\mu}\|} = \cos(\theta)$$

Note that

$$r = \cos(\theta) = \frac{\vec{x}^{\,\mu} \cdot \vec{y}^{\,\mu}}{\left\|\vec{x}^{\,\mu}\right\| \left\|\vec{y}^{\,\mu}\right\|} \approx \frac{\vec{x} \cdot \vec{y}}{\left\|\vec{x}\right\| \left\|\vec{y}\right\|} = \cos(\theta^*)$$

The mean of a footprint is normally close to zero since most footprint entries are close to zero so this is a reasonable approximation.

The normalized Euclidian distance (d_{norm}) is defined as:

$$d_{norm} = \left\|\vec{\chi} - \vec{\gamma}\right\|$$

where $\vec{\chi} = \frac{\vec{x}}{\left\|\vec{x}\right\|}$, and $\vec{\gamma} = \frac{\vec{y}}{\left\|\vec{y}\right\|}$

$$d_{norm} = \sqrt{2(1 - \cos(\theta^*))} \approx \sqrt{2(1 - r)}$$

There is a relationship between the cos(θ*) and normalized Euclidean (d_{norm}) by **Corollary 1** and because the angle between two vectors is the same as that between their unit vectors.

Therefore, the approximate relationship is demonstrated.