Mapping from Statistical to Biological Proximity

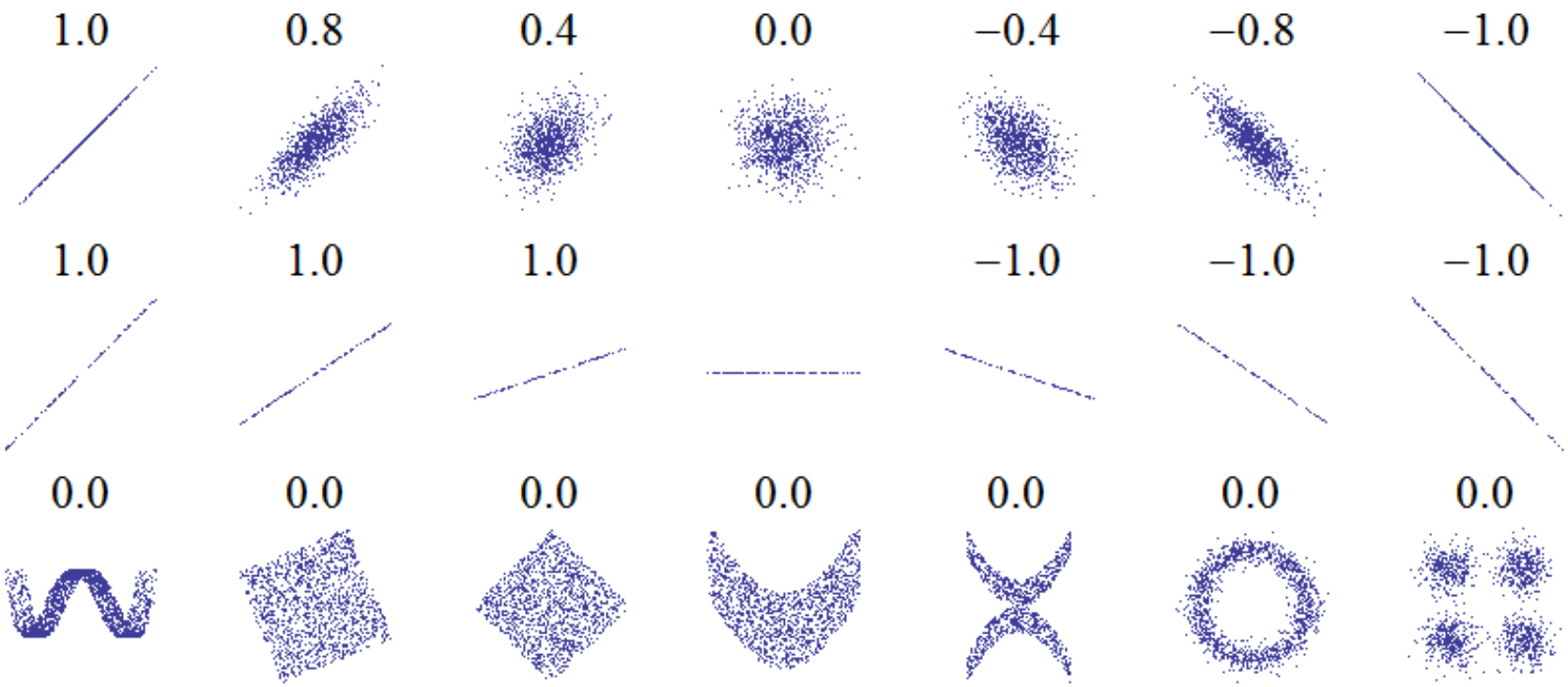
Malay Bhattacharyya

SRF, MIU, ISI Kolkata

Outline

- Statistical proximity
- Similarity/dissimilarity measures
- Gene ontology
- Biological proximity
- Dependence analysis
- Conclusions

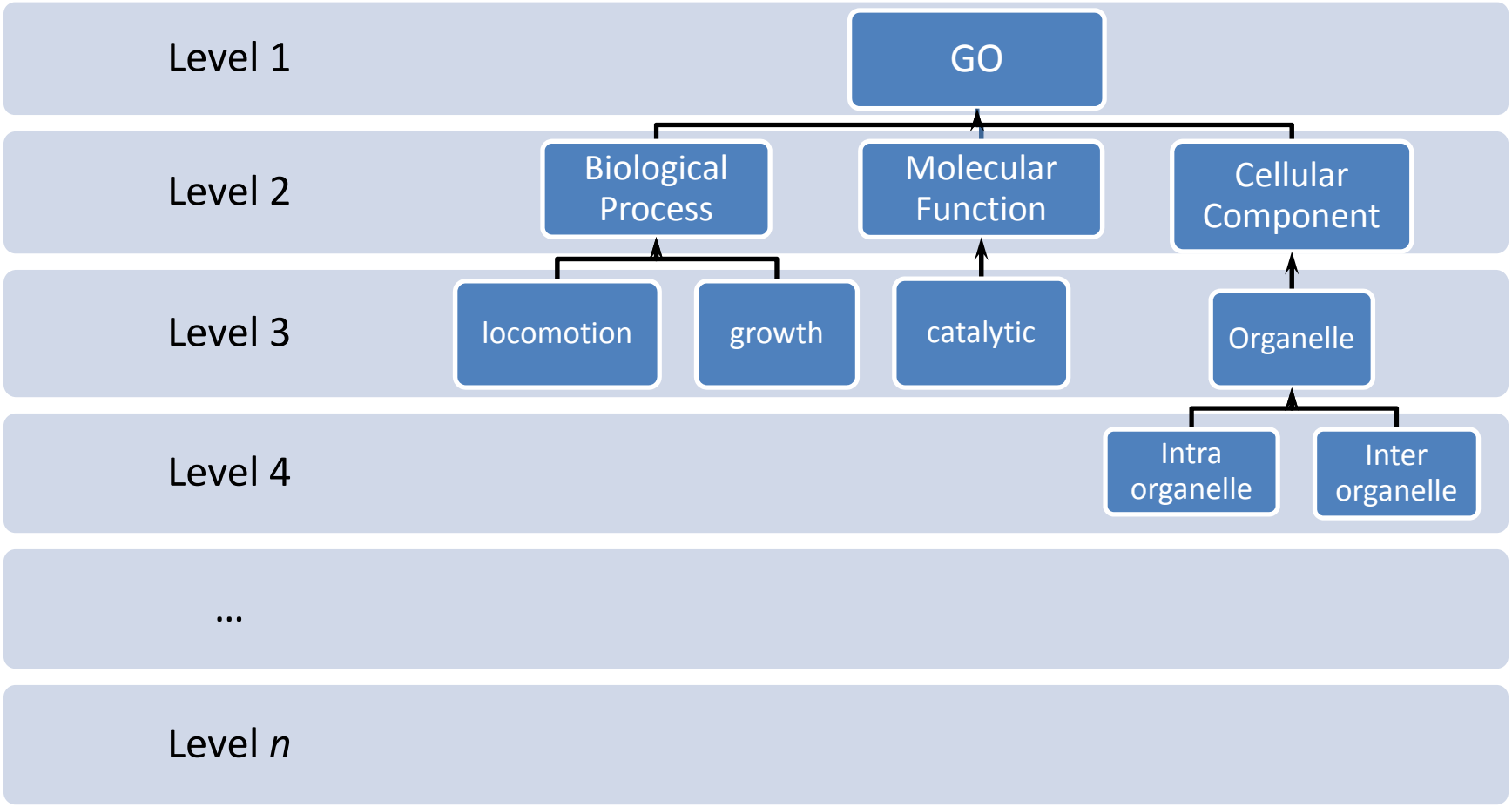
Statistical proximity



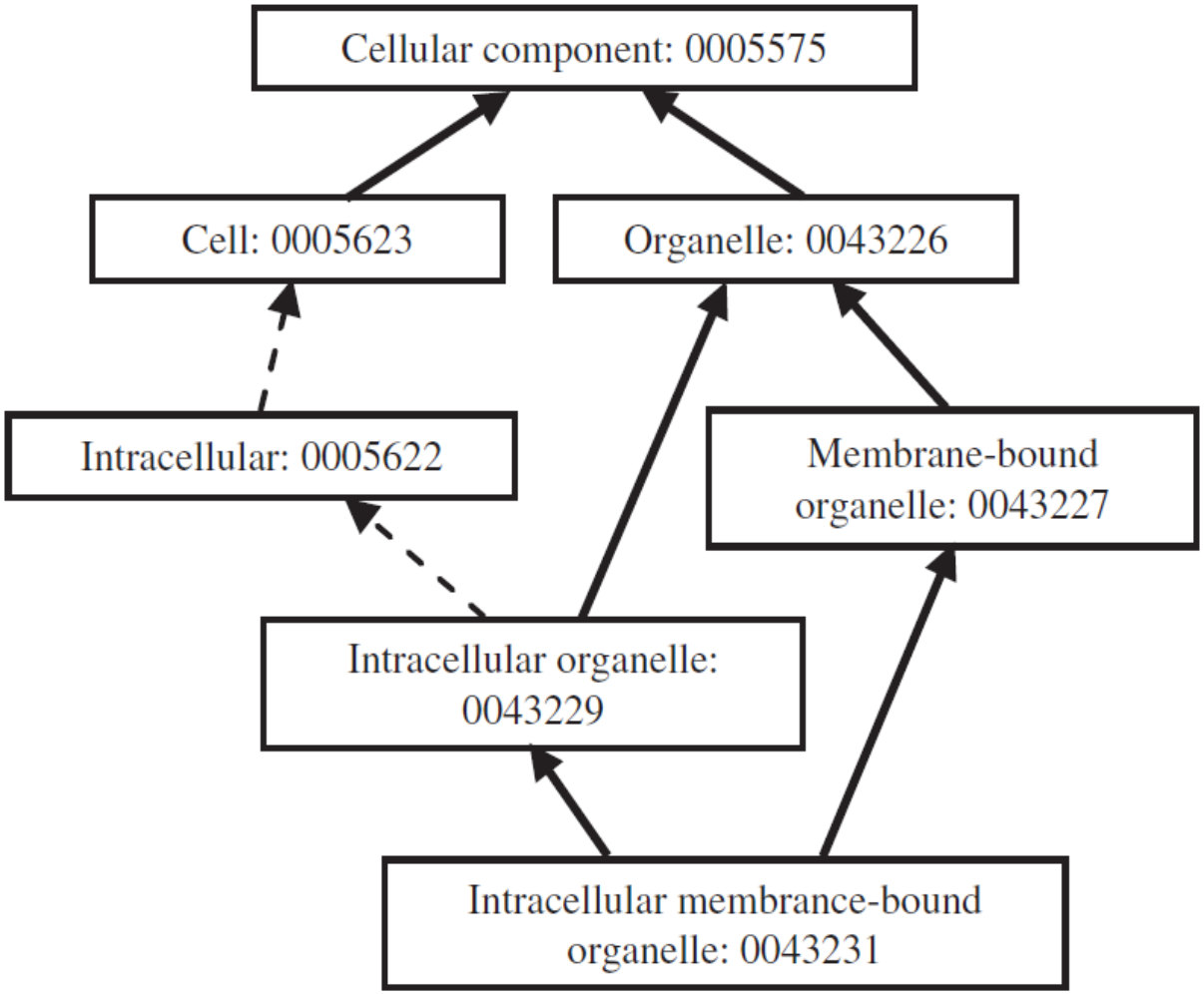
Similarity/dissimilarity measures

Name	Measure	Type
Uncentered correlation coefficient/Cosine	$(\mathbf{E}_i \bullet \mathbf{E}_j)/(\|\|\mathbf{E}_i\|\|\|\|\mathbf{E}_j\|\|)$	Similarity
Pearson correlation coefficient	$Cov(\mathbf{E}_i, \mathbf{E}_j)/(\sigma_{\mathbf{E}_i}\sigma_{\mathbf{E}_j})$	Similarity
Spearman's rank correlation	$\rho(Ranked(\mathbf{E}_i), Ranked(\mathbf{E}_j))$	Similarity
Cross-correlation 1	$\left(\frac{1-\rho(\mathbf{E}_i,\mathbf{E}_j)}{1+\rho(\mathbf{E}_i,\mathbf{E}_j)}\right)^{\beta}$	Distance
Cross-correlation 2	$\sqrt{2(1-\rho(\mathbf{E}_i,\mathbf{E}_j))}$	Distance
Root mean square	$\frac{1}{n}\sqrt{\|\|\mathbf{E}_i-\mathbf{E}_j\|\|^2}$	Distance
Minkowski	$\sqrt[p]{\|\|\mathbf{E}_i-\mathbf{E}_j\|\|^p}$	Distance
Squared Euclidean	$\|\|\mathbf{E}_i-\mathbf{E}_j\|\|^2$	Distance
City block/Manhattan	$\|\mathbf{E}_i-\mathbf{E}_j\|$	Distance
Chebyshev	$\max_t(\|\mathbf{E}_i(t)-\mathbf{E}_j(t)\|)$	Distance
Kullback-Leibler	$\sum_{t=1}^{n} e_j(t)\ln\frac{e_j(t)}{e_i(t)}$	Distance

Gene ontology



Gene ontology (a closer view)



Biological proximity [1]

Let us assume that for a pair of genes x and y the sets of annotated GO terms $T_x = \{t_{x1}, t_{x2}, \ldots, t_{xm}\}$ and $T_y = \{t_{y1}, t_{y2}, \ldots, t_{yn}\}$ are given, respectively. Then, the semantic similarity is computed as

- **Jaccard similarity:** $SS(x,y) = \frac{|x \cap y|}{|x \cap y|}$
- **Dice similarity:** $SS(x,y) = \frac{2|x \cap y|}{|x| + |y|}$

G-SESAME [2]

$$SS(x,y) = \frac{\sum_{1 \le i \le m} S(t_{xi}, T_y) + \sum_{1 \le j \le n} S(t_{yj}, T_x)}{m+n}$$

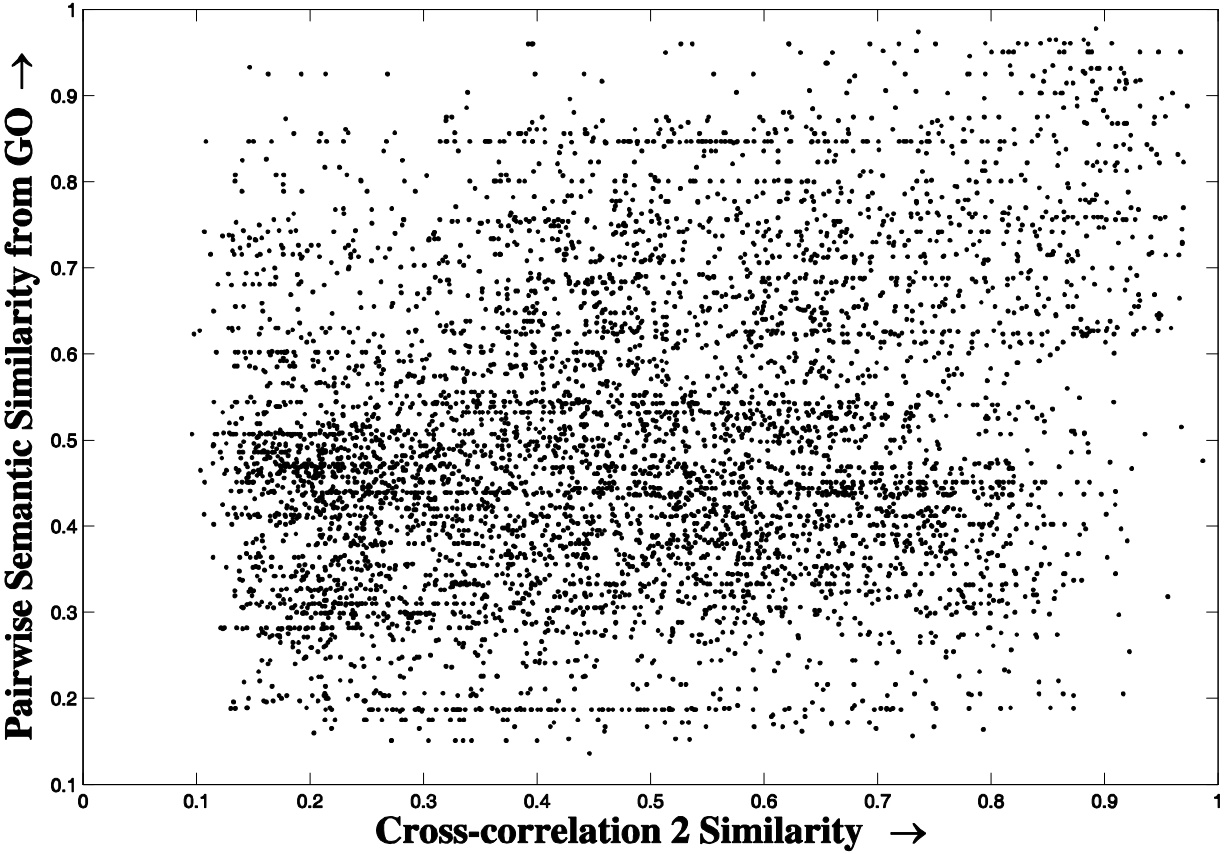
$$S(t,T) = \max_{t_i \in T} \frac{\sum_{tr \in \mathcal{T}_t \cap \mathcal{T}_{t_i}} (S_t(tr) + S_T(tr))}{\sum_{tr \in \mathcal{T}_t} S_t(tr) + \sum_{tr \in \mathcal{T}_{t_i}} S_{t_i}(tr))}$$

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e * S_A(t') | t' \in \text{childrenof}(t)\} \text{ if } t \neq A \end{cases}$$

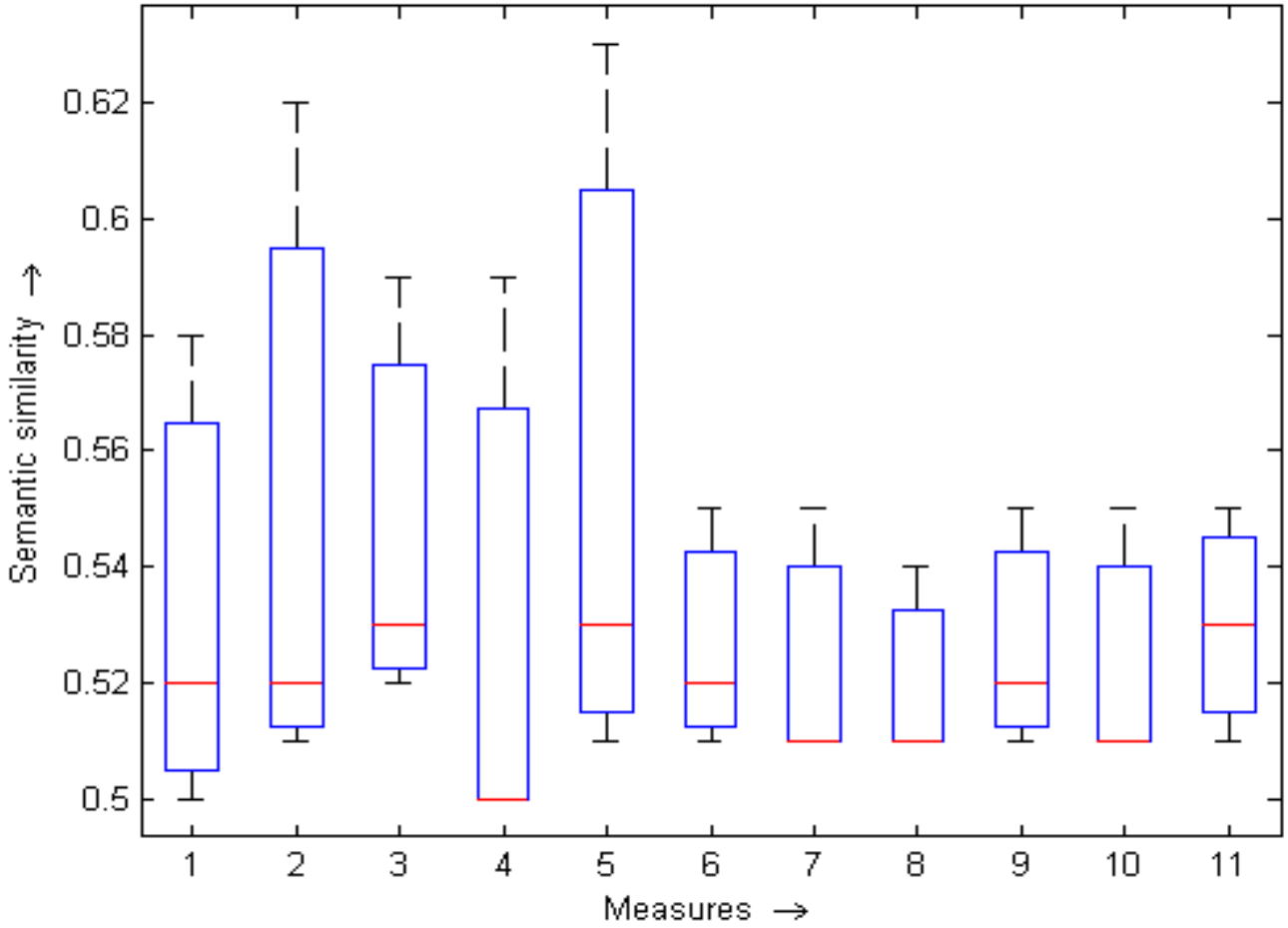
Dependence analysis

Similarity Measures	BP	CC	MF
Cosine	0.50	0.58	0.52
Pearson correlation	0.51	0.62	0.52
Spearman's rank correlation	**0.52**	0.59	**0.53**
Cross-correlation 1	0.50	0.59	0.50
Cross-correlation 2	0.51	**0.63**	**0.53**
Root mean square	0.51	0.55	0.52
Minkowski ($p = 3$)	0.51	0.55	0.51
Squared Euclidean	0.51	0.54	0.51
Manhattan/City block	0.51	0.55	0.52
Chebyshev	0.51	0.55	0.51
Kullback-Leibler	0.51	0.55	**0.53**

Dependence analysis (continued)



Dependence analysis (continued)



The *BioSim* measure [6]

$$BioSim = \frac{1}{n-1} \sum_{t=1}^{n-1} S(t)$$

$$S(t) = \frac{\alpha_1(t)\alpha_2(t)}{|\alpha_1(t).\alpha_2(t)|} . \frac{\cos(|\alpha_1(t)| - |\alpha_2(t)|)}{1 + \cos(\min(|\alpha_1(t)|, |\alpha_2(t)|)}$$

Results on *BioSim*

- For BP: 0.53
- For MF: 0.56
- For CC: 0.54

Biological evaluation of clusters

$$\mathcal{S}_{BP}(G) = \sum_{x,y \in G} SS(x,y),$$
$$\mathcal{S}_{MF}(G) = \sum_{x,y \in G} SS(x,y),$$
$$\text{and, } \mathcal{S}_{CC}(G) = \sum_{x,y \in G} SS(x,y), \text{ respectively.}$$

Conclusions

- More robust understanding of gene ontology
- More robust measure

References

1. P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, “Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation,” *Bioinformatics*, 19(10):1275-1283, 2003.
2. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C. -F. Chen, “A new method to measure the semantic similarity of GO terms,” *Bioinformatics*, 23(10):1274-1281, 2007.
3. H. Wang, F. Azuaje, O. Bodenreider and J. Dopazo, “Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships,” In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 25-31, 2004.
4. M. Popescu, J. M. Keller, and J. A. Mitchell, “Fuzzy measures on the gene ontology for gene product similarity,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3):263–274, 2006.
5. M. Mistry and P. Pavlidis, “Gene ontology term overlap as a measure of gene functional similarity,” *BMC Bioinformatics*, 9:327, 2008.
6. S. Bandyopadhyay and M. Bhattacharyya, “,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010 (in press).

Thank you