# Relating discrete annotation schemes in the functional space through literature analysis

M. Chagoyen[1], C.O.S. Sorzano[2], P. Carmona-Sáez[1], J.M. Carazo[1], A. Pascual-Montano[3]
[1]Biocomputing Unit. Centro Nacional de Biotecnología, CSIC. Spain
[2]Escuela Politecnica Superior, Universidad San Pablo-CEU. Spain
[3]Dpto. Arquitectura de Computadores. Universidad Complutense de Madrid. Spain

**Short abstract:**

We propose a methodology to create functional similarity measurements from data annotations using conceptual featural representations obtained from the analysis of relevant literatura. The literature contains our current state of knowledge regarding gene function. Therefore, it is a good source of data from which to establish functional associations.

**Objective:**

Establish a similarity measurement from functional annotations

**Abstract:**

Functional information associated with gene and gene products can be found in many biological databases. This information is usually provided by attachment of one or more annotations to a gene/gene product using a given controlled vocabulary or classification scheme. Although very valuable, current schemes present some conceptual limitations [1].

From the bioinformatics point of view, the first limitation is the shortage of automatic tools to compare and query data based on these functional annotations beyond exact match operations. To fill this gap, some authors have proposed functional similarity measurements based on the explicit relationships found in a given scheme (e.g. [2] makes use of GO ontology relations). However, these approaches present two shortcomings. First, associations not encoded in the ontology are not taken into account. Second, these measurements cannot be used when comparing data annotated with different annotation schemes.

The first limitation can be overcome by the enrichment of relationships in current functional schemes. Some examples are the use of association rules discovery of GO annotations in TIGR genomes [3], and linguistic relationships in GO categories [4].

In this work we propose a methodology to create functional similarity measurements from data annotations using conceptual featural representations obtained from the analysis of relevant literature. We postulate that the literature contains our current state of knowledge regarding gene function, and therefore it is a good source of data from which to establish functional associations.

Our approach is based on the Featural and Unitary Semantic Space hypothesis [5] that provides a model on how the meanings of object and action words are represented. Conceptual representations of functional categories are constructed by a set of terms frequently used in a corresponding literature corpus. Several similarity measurements are proposed on this featural space (similar to the document vector space used in document retrieval), or in derived latent semantic spaces obtained by SVD factorization.

**References:**

1. Fraser, A.G. and E.M. Marcotte, *A probabilistic view of gene function.* Nat Genet, 2004. **36**(6): p. 559-64.
2. Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
3. Kumar, A., A. Smith, and C. Borgelt. *Dependence relationships between Gene Ontology terms based on TIGR gene product annotations*. in *3rd International Workshop on Computational Terminology (CompuTerm 2004)*. 2004. Geneva.
4. Ogren, P.V., et al., *The compositional structure of Gene Ontology terms.* Pac Symp Biocomput, 2004: p. 214-25.
5. Vigliocco, G., et al., *Representing the meanings of object and action words: the featural and unitary semantic space hypothesis.* Cognit Psychol, 2004. **48**(4): p. 422-88.