Supplementary Text for “SimCT: a generic tool to visualize ontology based relationships”

1 Our notion of specificity: Precision of a term

The precision $p(t)$ of a term t in an ontology measures its specificity. Other measures have been proposed in the literature, like Information Content (IC) or Semantic Value (SV). Before giving the definition of precision, we shortly review these alternative definitions. In order to define precisely these measures, we begin with a brief glossary of the terms employed.

1.1 Glossary

- **Ancestor**: In an ontology, an *ancestor* of a node A is a node on a path from A to the root of the ontology;
- **Annotation**: An *annotation* is the association between a biological object and a term of the ontology considered; if an object is associated with more than one term we say that it has several annotations;
- **Biological object**: Any biological entity that has annotations to a biomedical ontology. For example genes, proteins, organisms, single nucleotide polymorphisms, etc...
- **Child**: A node A is the *child* of a node B if A is a descendant of B with a direct link between A and B;
- **Clustering tree**: The resulting tree of SimCT; its leaves are the (object|ontology term) pairs and the internal nodes are labeled with ontology terms and SRIs. This tree is a subgraph of the ontology;
- **Descendant**: A *descendant* of a node A is any node of which A is an ancestor;
- **Internal node**: A node of the clustering tree which is not a leaf;
- **Leaf**: A *leaf* is a node with no child;
- **Parent**: A node A is the *parent* of a node B if A is an ancestor of B with a direct link between A and B;
- **Path**: A *path* is a sequence of nodes such that from each of its nodes there is a direct link to the next node in the sequence;
- **Population**: The population of a term is the set of biological objects that are annotated to a term A directly, or to one of its descendants;
- **Precision**: Our measure of specificity for each term of an ontology;
- **Root**: A *root* is a node with no parent;
- **Similarity**: The similarity between two terms, or by extension between two (object|ontology term) pairs, is the precision of their most precise common ancestor in the ontology;
- **SRI** (Subtree Relevance Index): Numerical value helping to select interesting subtrees (see section 4 for the complete definition);

1.2 Information Content

Following Resnik [6], we can define the information content (IC) of a term t in an ontology as minus the logarithm of the probability of encounter, defined as follows: If we define $N_d(t)$ as the number of descendants of a term t (including t itself), and as N the total number of terms in the ontology, then the information content, normalized from 0 to 1 is given by:

$$IC(t) = -\frac{\log \frac{N_d(t)}{N}}{\log N}, \tag{1}$$

in which $N_d(t)/N$ represents the probability of occurrence of term t and its descendants. This definition has been extended for Gene Ontology in [5], where the probability of occurrence of a term t is related to the number of genes/proteins annotated to this term, or one of its descendants. If $P(t)$ represents the population of term t, and P the total number of genes/proteins annotated to this ontology, the information content is defined as:

$$IC_{Lord}(t) = -\frac{\log \frac{P(t)}{P}}{\log P}. \tag{2}$$

1.3 Semantic Value

Another measure of the specificity of a term in the ontology has been given in [7]. Called the *Semantic Value* (SV) it takes into account the semantic contribution of the ancestors of a term, giving different weights to the various relationships present in the ontology (*is a* or *part of* in the case of Gene Ontology). No information about the population i.e. the number of annotated genes/proteins is used here.

1.4 Precision

In SimCT, we have implemented a different measure of specificity which we call precision. While the information content as defined in [6] directly depends on the number of descendants (through the term N_d, see Eq. (1)), and the SV is the cumulative value of the semantic value of the ancestors (i.e. adding a *child term* to a term t does not change the semantic value $SV(t)$), we have combined both aspects in the definition of precision. The precision is based on the following ideas:

1. the more descendants a term t has, the less precise it is;
2. the more ancestors a term t has, the more precise it is.

Combining both ideas, we define the precision $p(t)$ as:

$$p(t) = -\frac{\log \frac{N_d(t)}{N \cdot N_a(t)}}{\log N \cdot N_a^{max}} \in [0, 1], \tag{3}$$

where $N_d(t)$ represents the number of descendants of t (counted only once), $N_a(t)$ the number of ancestors of t (counted only once) and N_a^{max} the maximal number of ancestors a leaf term can have in this ontology (corresponding to the "deepest" leaf).

1.5 Comparison of the different measures of specificity

We have computed the pairwise Spearman correlation coefficients between the four measures of specificity (IC, IC_{Lord}, SV and precision) applied on the GO sub-ontology Biological Process (see Table 1). Interestingly, our definition of precision has a high correlation coefficient with all three other measures, while all three others have a poor correlation coefficient with at least one of the other. This may indicate that our definition represents an interesting mixture of these different definitions.

Our definition has several interesting properties when compared to the three other measures:

- While the information content IC assigns a value of 1 to all leaves, our definition differentiates between leaves of short branches (lower precision) and leaves of long branches (higher precision). A support for the correctness of this is given when looking at the number of genes/proteins annotated to leaves in short or long branches in Gene Ontology. Clearly, long branch leaves (i.e. with

	IC	IC_{Lord}	SV	precision
IC	1	0.63	0.3	0.87
IC_{Lord}		1	0.23	0.6
SV			1	0.63
precision				1

Figure 1: *Pairwise Spearman rank correlation coefficient*

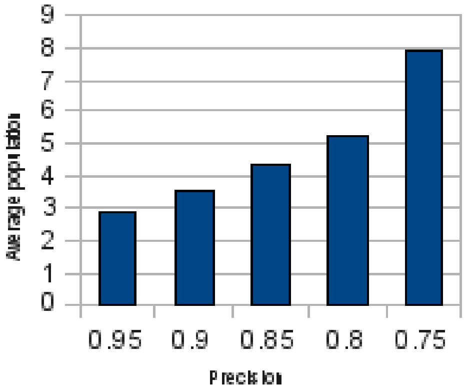


Figure 2: *Average number of proteins annotated to leave terms of the ontology as a function of the precision of these terms.*

	IC	IC_{Lord}	SV	precision
IC		36	27	30
IC_{Lord}			24	30
SV				30

Table 1: *Number of leaves that are identically classified between trees built using the various specificity measures. All trees have 41 leaves.*

higher precision) have on average lower populations than short branch leaves (with lower precision), indicating that their specificity is higher (see Fig. 2).

- Since the definition of precision (like IC and SV) does not take into account the population of the term, this definition can be applied to any ontology, allowing the inclusion of numerous ontologies into SimCT. Moreover, the precision is defined for all terms, unlike the IC_{Lord} based on population, that is ill-defined for terms with no genes annotated to (over 6000 terms in the GO sub-ontology Biological Process for example).

- In recent versions of Gene Ontology, new relationships have been added: *regulates*, *positively regulates* and *negatively regulates*. Moreover, other ontologies have other relationships (e.g. *connected to*, *preceded by,...*). Since the definition of SV requires to give a specific weight to different relationships, it is not obvious what the correct choice for these additional parameters would be for the various ontologies.

Despite the high degree of correlation between our measure of specificity (precision), and the three other measures (IC, IC_{Lord}, and semantic value SV), using one or the other measure in our clustering procedure would change the resulting tree slightly. We have compared the trees obtained using one of these four measures on an example, using a list of 25 human genes and the ontology biological process: *MTRR, ZNF175, LRRFIP2, EIF5B, HDAC9, ADRB2, NUAK2, TPST2, DLG5, KIR2DS4, KIR2DS1, KIR2DS5, KIR3DL2, KIR2DS2, KIR3DL1, SH2D2A, PARP8, CCL21, DUSP2, PTPN4, NCALD, PTGDR, MATK, KLRC3, KLRC2.*

The resulting trees have 41 leaves (due to multiple annotations of some proteins) and can be seen on Figure 3. The most relevant differences are the following:

- all four trees have in common the subtrees *"response to stimulus"*, *"organ development"* (or *"developmental process"*, which is a parent term of it) and *"cellular process"*.

- the trees based on IC and IC_{Lord} have subtrees *"immune system process"*, containing the proteins *HDAC9, CCL21, KIR2DS1, KIR2DS5, KIR2DS2, KIR3DL1.* For the 2 other trees (based on precision and SV), these proteins are located in the *"response to stimulus"* subtrees.

- the trees based on precision and IC_{Lord} both have subtrees annotated *"biological regulation"*, while the other two trees lack this subtree.
- the trees based on precision, IC and IC_{Lord} have subtrees annotated *"metabolic process"*, while the tree based on SV does not.

We have tabulated the number of leaves that are identically classified between the four different trees in Table 1. The numbers reflect the degree of correlation listed in Fig. 1.

The main advantages of precision compared to the three other measures are

- **simplicity**: it only depends on the structure of the ontology with no extra parameters;
- **generality**: it can be computed for any biomedical ontology.

2 Similarity

Once a measure of precision is given for all terms of an ontology, we have defined the similarity of two terms t1 and t2 as the precision of their most precise common ancestor. In SimCT, we do not define the similarity of two genes/proteins, but only the similarity of two terms, which is unambiguous.

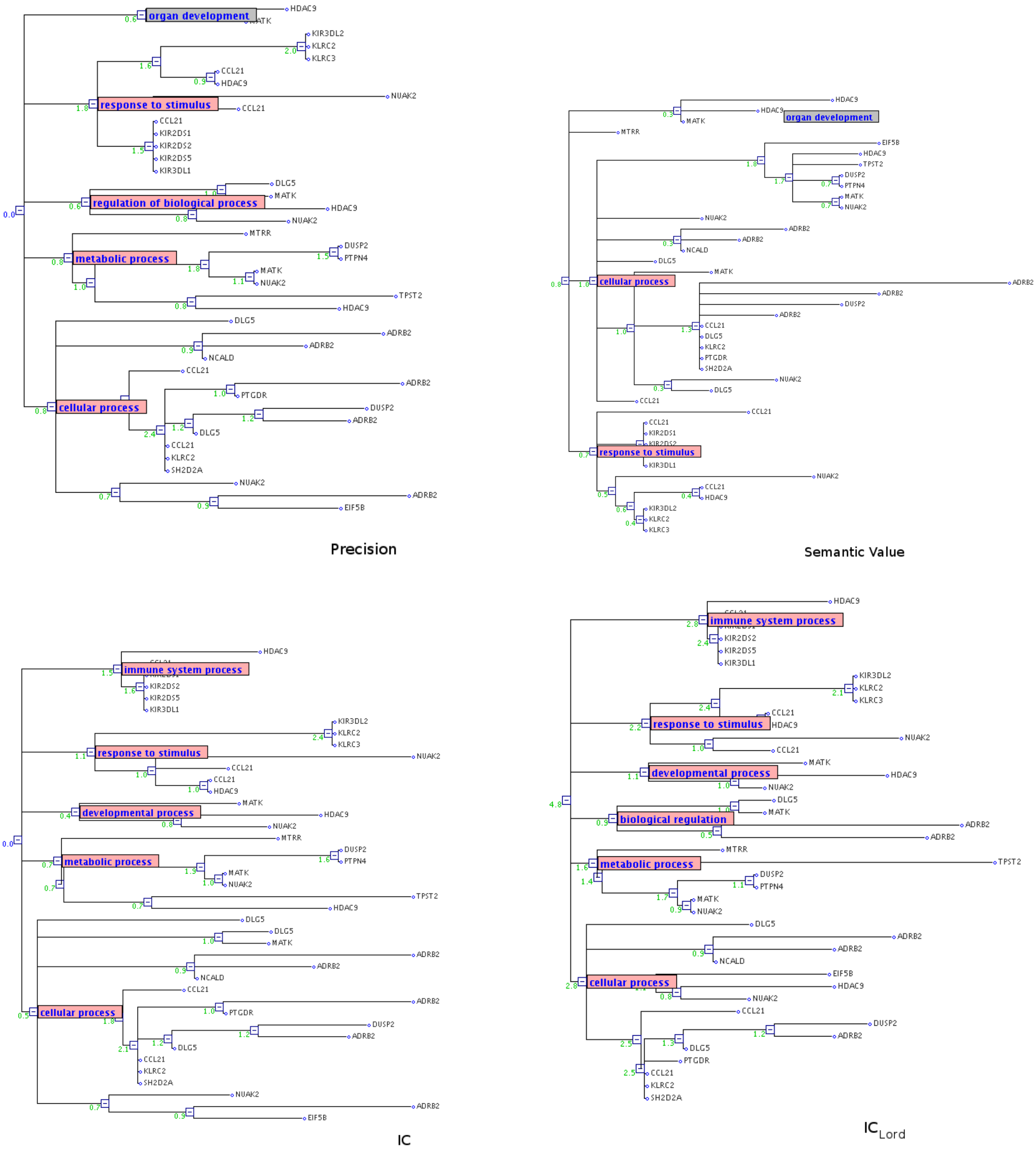
3 Clustering algorithm

We implemented a Kruskal-like algorithm for the hierarchical clustering, whose time complexity is in $O(n^2 \log n)$. This algorithm takes as input the set of all (objects|ontology term) pairs and progressively aggregates them in clusters according to their similarity (the more similar terms are aggregated first). The computation of the similarity between two clusters uses pre-computed results, like the precision of each term of the ontology, and the genealogy of terms. The order of the aggregation defines the final clustering tree. The algorithm is the following:

1. Build the initial list L_0 of clusters, consisting of all (object|ontology term) $(O_i|T_i)$: $L_0 = \{C_{i,0}\}$ where $C_{i,0} = (O_i|T_i)$. Initialize by computing the similarities between all pairs i, j of $C_{i,0}, C_{j,0}$; each pair (C_i, C_j) has a similarity $S_{i,j}$ and a most precise common ancestor (MPCA) $T_{i,j}$.
2. Rank the pairs of clusters by decreasing similarities; remove the top pair with similarity S_{top} and MPCA T_{top} from the list and add both cluster to the set of clusters that will be merged at step t+1: $C_{2Bmerged} = \{C_{k,t}, C_{l,t}\}$.
3. Scan the remaining list of pairs while $S_{i,j} = S_{top}$, and the $T_{i,j} = T_{top}$, and add the corresponding clusters to $C_{2Bmerged}$.
4. Clustering step:
   (a) remove all clusters in $C_{2Bmerged}$ from L_t;
   (b) add to L_t the newly formed cluster consisting of all members of $C_{2Bmerged}$, associated to term T_{top};
   (c) compute the similarity of the newly formed cluster to all remaining clusters of L_t;
5. Loop back to 2. until one final cluster remains.

4 Subtree Relevance Index (SRI)

Once the tree is built, we want to provide the user with a criteria to select the most relevant subtrees, i.e. the terms which are the most interesting to look at first. Numerous tools have implemented methods to evaluate the over/under-representation of Gene Ontology terms in a set of genes (see e.g. [3] for a review, or the Gene Ontology web site). All of these tools use the frequency of occurrence of a given term in a reference set (for example a full proteome) and compare it with the frequency of occurrence



Precision

Semantic Value

IC

IC_{Lord}

Figure 3: *Comparison of the four trees obtained using the four specificity measures discussed in the text. The annotations of the level 1 subtrees are indicated. Grey labels indicate that the term is the deepest available in this branch of the tree; in the opposite case, the label appears magenta.*

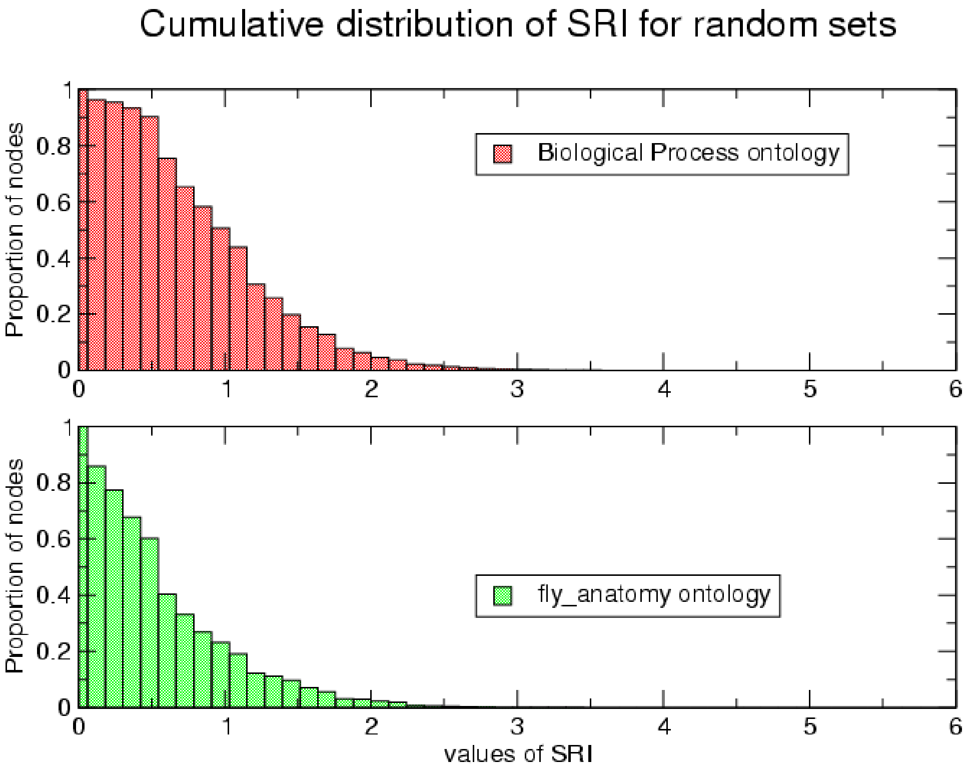


Figure 4: *Distributions of SRI values obtained from 500 random sets of drosophila proteins annotated to Biological Process and Fly Anatomy.*

in a test set, using various statistical measures. In SimCT, we wanted to develop a general method that can be applied to any ontology, so we don't use such measures since no systematic annotation resource is available apart from Gene Ontology. Instead, we have derived an empirical measure, which we call the Subtree Relevance Index (SRI), whose idea is the following: As we browse the clustering tree from the root to the leaves, the subtrees become smaller and smaller, and their associated terms more specific. Hence, the size of a subtree and the precision of its associated terms vary in opposite direction. As we want to highlight subtrees grouping many genes all annotated to a very specific terms, we define the SRI as follows :

$$SRI(T) = p(t) \times N(T), \tag{4}$$

where T represents a subtree, $N(T)$ its size (i.e. the number of leaves), t the term associated to the subtree and $p(t)$ the precision of this term. Empirical usage of this index shows that, in the case of Gene Ontology, subtrees with highest SRI often correspond to subgroups of genes with a statistically over-represented term.

In the Java visualization applet, the user can chose to open subtrees by decreasing value of the SRI, thus uncovering potentially most relevant subtrees first.

In order to evaluate the relevance of this index, and to estimate a threshold above which nodes are annotated to relevant terms, we simulated 500 random sets of 15 drosophila proteins, built the clustering trees, and recorded the values of the SRIs for the ontologies Biological Process and Fly anatomy. The cumulative distributions of the SRIs are plotted in Fig 4. Although these distributions depend on the ontology considered, we can estimate that nodes with SRIs above 2.5 can be considered as representing relevant annotations.

5 Comparison with other tools

5.1 DAVID and GO::TermFinder

As described in the main text, we compared the SRI values and the p-values computed by the DAVID software [2] and the GO::TermFinder implementation [1] for a set of genes related to natural killer cells in human (this is the set of genes used for the demo on the SimCT web-site). Table 2 presents the results.

GO IDs	GO terms	SimCT	SimCT	DAVID	DAVID	GO::TermFinder
		SRI	raw pval.	raw pval	corr. pval	corr. pval
GO:0006952	defense response	3.3	1×10^{-5}	3.7×10^{-7}	1.9×10^{-3}	1.2×10^{-7}
GO:0006968	cellular defense response	3.4	1×10^{-6}	5.3×10^{-6}	6.9×10^{-3}	3.8×10^{-5}
GO:0007154	cell communication	5.3	5×10^{-3}	5.1×10^{-3}	9.3×10^{-1}	7.7×10^{-8}
GO:0007165	signal transduction	6.8	3.1×10^{-3}	3.5×10^{-3}	8.7×10^{-1}	6.3×10^{-7}
GO:0007166	cell surf. recept. sig. trans.	4.6	3×10^{-3}	4.5×10^{-2}	1	3.3×10^{-4}
GO:0034961	cellular biopol. biosynt. process	2.6	0.7			

Table 2: *Comparison of the SRI and the p-values for the most relevant nodes of the SimCT tree.*

It is surprising to note that DAVID and GO::TermFinder give completely different values for the raw p-values. This might be due to the fact that they use different reference sets or different versions of annotation files. In SimCT, we compute the raw p-value using a hypergeometric distribution, taking as reference set the complete annotated proteome. We take into account whether the user has chosen to ignore electronic (IEA) annotations. We note that in this example, the term with highest SRI (*"signal transduction"*) is not the most statistically over-represented term. However, it is a highly *represented* and *specific* term, with 21 out of 53 proteins annotated to it.

5.2 GOSurfer and GOTreePlus

GOSurfer [1] and GOTreePlus [4] are two tools that can be compared to SimCT: the user inputs a gene list, and the tool displays a portion of the ontology that is most relevant to the functions represented by these genes. Both tools represent a portion of the DAG as a tree (like SimCT does), but to the best of our knowledge, neither publication gives details regarding the method used to perform this transformation, like we do for SimCT. Another difference is the visualization method: GOTreePlus is based on the .NET programming environment, providing a very smooth navigation through the ontology, whereas GOSurfer uses a more static visualization, in which the tree cannot be edited or modified. Both tools use statistical criteria to highlight over/under-represented terms in the list. The main difference to SimCT are that these tools are standalone applications which must be downloaded and installed locally to run, and that they are restricted to the Windows operating system. There is no such limitation in SimCT, which is a web-based application which merely requires Java to be installed. Moreover, SimCT can be used for any biomedical ontology among those supported (25 currently), and is not restricted to Gene Ontology as are both tools.

References

[1] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–5, Dec 2004.

[2] Glynn Jr Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[3] Purvesh Khatri and Sorin Drghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–95, Sep 2005.

[4] Bongshin Lee, Kristy Brown, Yetrib Hathout, and Jinwook Seo. Gotreeplus: An interactive gene ontology browser for proteomics projects. *Bioinformatics*, Feb 2008.

[5] 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):1275–83, Jul 2003.

[6] P Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[7] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–81, May 2007.