GOTEA in VisANT

The four steps here describe how GOTEA works in VisANT. For illustration purposes, the following steps only take one metanode, *G*, into account and only calculate the enrichment score of one target GO term, *T*.

Step 1: Fully annotate all of the nodes in *G* with gene names and GO terms.

Step 2: Calculate density scores for each node based upon the topology and the GO term similarity to *T*. A vector D^G of density scores of each gene in *G* is computed, with the element of D^G for the *i*th gene denoted D_i. The density score is used to evaluate the impact of other genes in *G* on the *i*th gene, according to both the GO term similarity and the topological distance to the *i*th gene. D_i is defined as:

$$D_i = \sum_{j \in G} \log_2 \left[\left(\frac{M_j}{\alpha} \right) \Theta\left(M_j - \alpha\right) + \Theta\left(\alpha - M_j\right) \right] \mathrm{e}^{-\beta d_{ij}},$$

where the step function,

$$\Theta(x-y) = \begin{cases} 1 & x \geq y \\ 0 & x < y, \end{cases}$$

ensures that $D_i \geq 0$. M_j is a measure of the GO term similarity calculated based upon the graph structure of the GO term hierarchy (Wang, Du et al. 2007). A significance threshold, α, is used to control the contribution that gene *j* makes to D_i. For larger α, a greater number of less statistically significant (with $M_j < \alpha$) genes are filtered and do not contribute to D_i. The shortest distance between genes *i* and *j* given the topology of *G* is denoted d_{ij} and was calculated with the Floyd-Warshall algorithm (Floyd 1962). We assume that shorter distances make an exponentially greater contribution to the density than do longer distances, with the steepness of the exponential determined by the parameter β. When a bigger β is chosen, more distant genes can contribute to the density. Taken together, the parameters α and β are used to control the sensitivity and selectivity of the density.

Step 3: Another vector of density scores, D^{NG}, is computed based upon a randomly chosen subset of genes representative of the background distribution. The background consists of all genes annotated by NCBI.

Step 4: Statistical significance for rejecting the null hypothesis is determined by a permutation test. For statistical robustness, step 3 is repeated n times. The number of times the average density score of randomly chosen genes is found to be larger than the average density score of genes in *G* is counted after n iterations and used to compute the final p-value.

These 4 steps can be carried out for multiple testing by using multiple metanodes and multiple targeting GO terms. In this case, the p-values are corrected using FDR methods (Benjamini, Drai et al. 2001). Specifically, $FDR = p \times m/k$, where *m* is the total number of GO terms tested and *k* is the rank of the GO terms under consideration. There is also an option for

GOTEA to identify representative GO terms from all its discoveries based upon approaches that identify the most informative GO term (Zhou, Kao et al. 2002).

NMEA in VisANT

NMEA are implemented in a manner similar to GOTEA. Where GOTEA used GO term similarities, NMEA uses p-values from T-tests on the expression values of two phenotypes.

Step 1: Fetch the expression profile of each gene in a given module (i.e. metanode, denoted *M* in the following context) from formatted user input. The input should include an adequate number of samples with comparable phenotypes (e.g. normal and disease).

Step 2: A vector D^M of density scores of each gene is computed, with the element of D^M for the *i*th gene denoted as D_i. D_i is defined as:

$$D_i = \sum_{j \in G} \log_2 \left[\left(\frac{\alpha}{M_j} \right) \Theta\left(\alpha - M_j\right) + \Theta\left(M_j - \alpha\right) \right] \mathrm{e}^{-\beta d_{ij}},$$

where the step function,

$$\Theta(x-y) = \begin{cases} 1 & x \geq y \\ 0 & x < y, \end{cases}$$

ensures that $D_i \geq 0$. M_j is the *p*-value from a two-tailed t-test of differential expression between two phenotypes (for example, normal and disease). The parameters α and β are used to control the sensitivity and selectivity of the density as described in the previous section.

The density score is used to evaluate the impact of other genes in *M* on the *i*th gene, according to both the p-value calculated by T-test (an indicator of differential expression) and their topological distances to the *i*th gene.

Step 3: Another vector of density scores, D^{NM}, is computed by randomly shuffling the phenotypes to obtain a representative sampling of the background distribution.

Step 4: Statistical significance for rejecting the null hypothesis is determined by a permutation test. For statistical robustness, step 3 is repeated n times. The number of times the average density score of randomly chosen genes is found to be larger than the average density score of genes in *M* is counted after n iterations and used to compute the final p-value.

When applying NMEA to multiple metanodes, the p-value must be corrected by FDR in a manner similar to what was described above for GOTEA. In this case, $FDR = p \times m/k$ as before, but *m* is the total number of metanodes and *k* is the rank of the metanodes under consideration.

References

Benjamini, Y., D. Drai, et al. (2001). "Controlling the false discovery rate in behavior genetics research." Behav Brain Res **125**(1-2): 279-84.

Floyd, R. W. (1962). "Algorithm 97: Shortest path." Commun. ACM **5**(6): 345.

Wang, J. Z., Z. Du, et al. (2007). "A new method to measure the semantic similarity of GO terms." Bioinformatics **23**(10): 1274-81.

Zhou, X., M. C. Kao, et al. (2002). "Transitive functional annotation by shortest-path analysis of gene expression data." Proc Natl Acad Sci U S A **99**(20): 12783-8.

Tables

Table S1. Ranked list of all informative GO terms detected in the KEGG human cell cycle pathway.

Category	ACC	Term name
Biological	7049	cell cycle
Process	6259	DNA metabolic process
	48523	negative regulation of cellular process
	48522	positive regulation of cellular process
	6355	regulation of transcription, DNA-dependent
	42981	regulation of apoptosis
	50790	regulation of catalytic activity
	6366	transcription from RNA polymerase II promoter
	42127	regulation of cell proliferation
	6468	protein amino acid phosphorylation
	51704	multi-organism process
	6996	organelle organization
	48513	organ development
	9605	response to external stimulus
	42221	response to chemical stimulus
	6511	ubiquitin-dependent protein catabolic process
	3	reproduction
	42592	homeostatic process
	7399	nervous system development
	34621	cellular macromolecular complex subunit organization
	65003	macromolecular complex assembly
	9966	regulation of signal transduction
	48468	cell development
Cellular	43234	protein complex
Component	5654	nucleoplasm
	44430	cytoskeletal part

	5730	nucleolus
	5789	endoplasmic reticulum membrane
	44421	extracellular region part
	31982	vesicle
	5829	cytosol
	31967	organelle envelope
	5794	Golgi apparatus
	44429	mitochondrial part
	30529	ribonucleoprotein complex
Molecular Function	4672	protein kinase activity
	30234	enzyme regulator activity
	5102	receptor binding
	5524	ATP binding
	17111	nucleoside-triphosphatase activity