Exploiting statistical methodologies and controlled vocabularies for prioritized functional analysis of genomic experiments: the StRAnGER web application

Aristotelis A. Chatziioannou* and Panagiotis Moulos

Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece

StRAnGER is a web application for the automated statistical analysis of annotated gene profiling experiments, exploiting controlled biological vocabularies, like the Gene Ontology or the KEGG pathways terms. Starting from annotated lists of differentially expressed genes and gene enrichment scores, regarding the terms of each vocabulary, StRAnGER repartitions and reorder the initial distribution of terms to define a new distribution of elements. Each element pools terms holding the same enrichment score. The new distribution thus derived, is reordered in a decreasing order to the right, according to the observation score of the elements, while elements with the same score, are sorted again in a decreasing order of their enrichment scores. By applying bootstrapping techniques, a corrected measure of the statistical significance of these elements is derived, which enables the selection of terms mapped to these elements, unambiguously associated with respective significant gene sets. The selected terms are immunized against the bias infiltrating statistical enrichment analyses, producing technically very high statistical scores, due to the finite nature of the data population. Besides their high statistical score, another selection criterion for the terms is the number of their members, something that incurs a biological prioritization in line with a Systems Biology context. The output derived, represents a detailed ranked list of significant terms, which constitute a starting point for further functional analysis.

Keywords: ontological analysis, pathway analysis, gene enrichment analysis

INTRODUCTION

DNA microarrays (cDNA or oligonucleotide) constitute a widely used measuring technology regarding the response of whole genomes, yielding estimates of total gene expression in selected species. Measuring the comparative gene expression with microarrays, has a critical importance in the analysis of biological control mechanisms, phenotyping, molecular profiling of diseases, and more accurate disease classification. Applications include the classification of tumors with different prognosis, indistinguishable solely by microscopic examination or optical inspection, derivation of diagnostic or prognostic signatures, monitoring responses to multiple or different treatments or therapies, rational drug design (Maynard et al., 2003; Tarca et al., 2006). In general, DNA microarrays promote the understanding of the plasticity of cellular circuitry and how this is linked to the manifestation of a phenotypic versatility, regarding various cellular functions in different organisms.

Today, several software packages, commercial or open source, perform routinely microarray analysis and interpretation usually deriving lists of differentially expressed (DE) genes with several hundreds of genes. Though indispensable, the derivation of DE gene lists still remains a primary step, failing to provide insight on the underlying molecular mechanisms, governing the biological problem interrogated and leaving the experts bewildered. In order to highlight statistically significant and biologically relevant actors, enabling thus a systemic perspective, a new round of analysis is applied, emphasizing in the involvement of molecular pathways rather than isolated genes, in order to uncover sets of genes that participate or regulate the same cellular pathway. For this scope, biological ontologies, or more general controlled biological vocabularies constitute valuable sources of standardized biological information, appropriate for DNA microarrays functional analysis. The Gene Ontology (GO; Ashburner et al., 2000) provides such functional annotation adopting a hierarchical schema. In addition, the Kyoto encyclopedia of genes and genomes (KEGG) biological pathway database (Kanehisa et al., 2010) comprises a well structured and constantly enriched library of molecular networks, which has been widely used as a reference point for biological interpretation of large-scale datasets.

Nowadays, multiple software tools are targeting the issue of detecting over-represented ontological terms, in processed biological datasets (Dennis et al., 2003; Hosack et al., 2003; Beissbarth and Speed, 2004; Boyle et al., 2004; Martin et al., 2004; Zhong et al., 2004; Conesa et al., 2005; Zhang et al., 2005; Bauer et al., 2008; Zheng and Wang, 2008). The majority are statistical implementations of tests, which estimate the number of successes in a sequence of draws from a finite population without replacement, like the hypergeometric distribution. Generally, the detection of an over-represented ontological term (e.g., a GO term) can be accurately approximated by the well known example of drawing two-colored balls from an urn, where the balls are equal to the number of genes. Testing statistically an ontological term corresponds to drawing the genes linked to it and examining which of them are parts of the DE list and which are not, in other words...
counting white (members) and black (not-members). This process leads to the derivation of a contingency table with four categories for each ontological term: (i) DE genes annotated at this term, (ii) DE genes not annotated at this term, (iii) non-DE genes annotated at this term, and (iv) non-DE genes not annotated at this term. A proper statistical test is then used to determine the extent of over-representation of a term in the DE gene list, compared to a reference set. As the number of the ontological terms in each dataset may vary from several tens to hundreds or even thousands of terms, the probability of false positives rises, and this requires application of several, possible, multiple correction methodologies (Osier et al., 2004). Apart from gene names, other software can accommodate DNA sequences (Conesa et al., 2005) in order to perform the aforementioned analysis.

There are various software implementations of the aforementioned approach, each with its own advantages and limitations. A number of them (GO-TermFinder, GoSurfer, Blast2GO, EASE; Hosack et al., 2003; Boyle et al., 2004; Zhong et al., 2004; Conesa et al., 2005) supports only one statistical mode for the detection of over-represented ontological terms, while others present limitations in their capability to support various available microarray platforms, experiment types, or multiple organisms (Dennis et al., 2003; Beissbarth and Speed, 2004; Boyle et al., 2004; Martin et al., 2004). Some, like GO-TermFinder for instance, require from the user to provide a suitable set of genes, which is used as reference set for the statistical enrichment test (Boyle et al., 2004), while at the same time, certain programming skills are required from the user, in order to exploit the software. Others, being user friendly, require only specific gene accessions but with the cost of becoming inflexible in integrating further relevant information (i.e., gene specific p-values, gene expression values or other statistics) to their output. Moreover, certain tools are inflexible regarding data import, requiring often time consuming manipulations by the user like the Ontologizer (Bauer et al., 2008), while DAVID results in complex outputs (Dennis et al., 2003), bewildering the researcher with a deluge of statistical measures, that confound the interpretation procedure.

The intelligible analysis and interpretation of ontological terms is inextricably bound to the implementation of efficient visualization mechanisms, which illustrate the functional relations among over-represented terms, as well as with all of the terms, neatly and hierarchically. Although various tools provide a visual representation of the results (usually in tree-like demonstrations), still many, like GOToolBox for instance, lack either the feature of automated generation or the capability of visual representation (Martin et al., 2004). Finally, a different category of tools, exploiting ontologies but giving emphasis around specific gene sets, utilize the entire gene lists of a given high-throughput experiment, after sorting them in terms of expression, rather than smaller lists of selected, based on statistical or empirical thresholds, entities. Their algorithms estimate the enrichment of their top or bottom ranked genes to the aforementioned, particular gene sets, as they are formed based on their functional annotations, utilizing different statistical scores coupled with resampling techniques for validation. Typical tools of this category are GSEA (Subramanian et al., 2005) and ErmineJ (Lee et al., 2005). These tools seem suitable for pairwise experimental designs, for example control vs treatment or disease configurations (Huang et al., 2009), a feature which proves to be limiting, if more complex experimental designs, such as time-course studies or multiple drug–response experiments, are targeted.

A common inherent problem, typical of the statistical methodologies that test over-representation, stems from the hierarchical structure of ontologies and their finite nature. Thus, many terms describing biochemically very particular sub-functions of a given cellular process, are ranked very high, simply because they are linked to very small gene numbers. This trivial finding from a pathway perspective, results from the fact that such terms possess a very strong statistical score, since their enrichment is very high or even complete. Such terms represent at the same time a large proportion of the total annotations of a dataset obfuscating the interpretation, with descriptions of limited information content from a pathway perspective. Multiple-testing correction methods do not remedy the problem, as their properties have not yet been adequately studied (Osier et al., 2004) and most of them do not take into account the graph structure of biomedical ontologies (Goeman and Mansmann, 2008). As high-throughput technologies are noise sensitive, the presence of false positives, due to technological limitations, may severely contaminate the interpretation. Processes corresponding to these GO terms (or more general similar, controlled vocabularies) may thus have limited functional biological value. Graph theoretic approaches are addressing this issue by considering the tree structure of the description. For example, GOToolBox (Martin et al., 2004) uses a combination of modules that detect over-represented GO terms, clusters them and scores genes based on their shared GO terms using a Czekanowski-Dice like distance, while Ontologizer (Bauer et al., 2008) considers the GO inheritance problem that is, the fact that the probability of a GO term being significant is much higher if one or more of its parent terms are significant. Ontologizer addresses this issue by implementing the parent–child method (Grossmann et al., 2007) or the elim algorithm (Alexa et al., 2006). However, both tools have certain particularities regarding data import, which limit their applicability.

StrAnGER is a web-based application1, which performs functional analysis of high-throughput -omic datasets, on an initial list of significant entities, derived after applying statistical, and/or empirical thresholds. At present, it utilizes GO or the KEGG pathway database. However, the implementation is generic enough to accommodate other available biological ontologies, related to -omic data or more generally controlled vocabularies. StrAnGER uses established statistical methods, in order to relate the identified significant genes with important nodes in the GO tree structure or interchangeably map those genes to over-represented metabolic pathways, like KEGG Pathways. Aim of StrAnGER is to suggest whole molecular pathways or parts of them, incorporating a crucial number of significantly DE genes of the list as interesting targets for further biological research rather than isolated genes, which are more susceptible to the impact of systematic or random errors. The main goal of StrAnGER is to sort out among all terms of a controlled vocabulary associated with the significant gene list, those revealing critical aspects of cell function (i.e., nodes higher in the GO hierarchy, densely enriched KEGG metabolic pathways), which consequently encompass a reliably high number of genes that implies unambiguously the involvement of a specific biochemical pathway. These terms are ranked according to their statistical significance, following their p-value score as derived from a suitable over-representation test. In this sense, the impact of noise on high-throughput genomomic experiments is significantly mitigated, thus consolidating the selection of specific biological targets, for further investigation.

1http://www.grissom.gr/stranger/
**MATERIALS AND METHODS**

**DATA IMPORT AND SUPPORTED BACKGROUND SETS**

StRAnGER requires a tab-delimited text file, with unique gene identifiers corresponding to the microarray platform used or to the public database that is chosen for genomic annotation extraction (e.g., Ensembl). An additional column, with p-values corresponding to each gene, is optional. StRAnGER offers the possibility to use a variety of sources and organism-related information, for the generation of the background dataset, including Bioconductor (Gentleman et al., 2004) array annotation packages or Ensembl gene annotations for various organisms (Flieck et al., 2008). In addition, the user can upload a custom annotation file, including the minimum information required for StRAnGER analysis, in tab-delimited text format. In all cases, a user-friendly wizard enables the user to specify the number and type of columns, containing information needed for the subsequent analysis.

**STATISTICAL ASSIGNMENT OF ENRICHED ONTOLOGICAL TERMS**

StRAnGER currently provides three statistical tests for the identification of enriched ontological terms, given a list of selected entities (i.e., genes, but also proteins or other molecules), and together, the appropriate reference set. If \( t \) is the total number of significant entities, \( z \) the number of significant entities related to a term \( T_i \), \( n \) the total number of the members of the reference set, and \( x \) the members of the reference set related to a term \( T_i \), then

1. The hypergeometric test, where the probability for a term \( T_i \) to be over-represented is given by the formula:

\[
P_{hi}^z (Z = z > s,n,t,x) = \sum_{i=1}^{t} \binom{t}{z} \binom{n-z}{x-z} / \binom{n}{x},
\]

2. The Fisher’s exact test where the probability for a term \( T_i \) to be over-represented is given by the formula (directly related to the hypergeometric distribution):

\[
P_f^z (z,n,t,x) = \binom{z+x}{z} \binom{t-z+n-x}{t-z} / \binom{n}{t},
\]

3. The \( \chi^2 \) test, where the \( \chi^2 \) statistic for a term \( T_i \) is calculated firstly as follows:

\[
A = \frac{(z+x)t}{t+n}, \quad B = \frac{(z+x)n}{t+n}, \quad C = \frac{(t-z+n-x)t}{t+n}, \quad D = \frac{(t-z+n-x)n}{t+n}.
\]

Then the \( \chi^2 \) statistic is given by

\[
\chi^2 = \frac{(A-B)^2}{A} + \frac{(B-C)^2}{B} + \frac{(C-D)^2}{C} + \frac{(D-A)^2}{D}
\]

And the probability for \( T_i \) to be over-represented is given by (assuming \( w = \chi^2 \) and one degree of freedom for the \( \chi^2 \) cumulative distribution function)

\[
Pr^T_i (X > w, 1) = \gamma\left(\frac{1}{2}, \frac{w}{2}\right) / \Gamma\left(\frac{1}{2}\right),
\]

where \( \gamma \) denotes the lower incomplete Gamma function and \( \Gamma \) denotes the Gamma function (Soong, 2004).

**THE StRAnGER ALGORITHM**

The lists of statistically significant entities obtained after the high-throughput data processing are utilized to perform functional analysis, exploiting ontological terms. Each ontological term has a certain score (number of observations) both in the whole array list just as in the significant list. The ratio of the number of observations of a certain term in the significant list to its number of observations in the whole array or more general reference list is called “enrichment” and is considered a feature. Having defined the enrichment objects, the significance of a term is assessed through the use of a proper statistical test. Appropriate statistical tests for this analysis are the aforementioned hypergeometric test, the \( \chi^2 \) test or the Fisher exact test.

At this point, the need to identify the desirable elements from the biological point of view should be stressed, namely, terms with a low p-value score, that comprise a substantial number of genes linked to this term in the whole reference list, which at the same time present high enrichment. In the case of GO terms, the GO schema defines a hierarchical tree for the representation of genes, distinguishing description of gene in three principal categories, viz., biological process, molecular function, cellular component. All complex biological actions within a cell are perceived by adopting a “top-down” logic, as functional entities, which can be analyzed to a combination of cellular, biochemical (metabolic, signaling, transport, etc.) pathways. Every process can be broken down to numerous other elementary processes, linked with subsets of the parent gene set, to the point of trivial cases of GO terms, related to only one or two genes, presenting therefore an extremely high enrichment of (50–100%), which is reflected in a very low p-value.

The processes corresponding to these GO terms have limited biological value from a global, cellular perspective. Nevertheless, they have a very strong statistical score merely because their enrichment is 100%. To overcome this problem, and define a set of ranked terms with both high biological content and statistical significance, the following steps are adopted.

Many terms have identical ratios (enrichments). By grouping terms according to their specific enrichments, a distribution of elements is created, which pools together terms with the same enrichment, and is then sorted from the most to the less frequent cases (left to right order). Elements with the same frequency are sorted again in a decreasing order, according their p-value enrichment scores. Using this list, a figure depicting the relative frequency of each group is created (Figure 1A).

In this way, elements are sorted by taking into account both the statistical and biological significance and defining a final population distribution for the ranked elements. In this distribution, a percentile threshold is set (default 90th percentile) to define the acceptable cutoff for significant terms. By applying bootstrapping in the thus formed distribution (Efron and Tibshirani, 1993), a corrected assessment of the statistical significance of
these elements is enabled, which targets the selective gleaning of terms mapped to these elements. These terms are unambiguously associated with respective sets of differentially expressed genes. Using the population of the elements as the starting sample pool, a number of bootstrap iterations (default 10000) is performed. Each bootstrap yields a new element distribution. From the frequencies of each element distribution, a histogram is built and a cutoff threshold is derived. Based on all these cutoff thresholds, a distribution of the ranked thresholds is built, which is compared with the cutoff threshold of the initial element distribution. An illustrative description of the distribution of the cutoff thresholds of the bootstrap technique in the present analysis is given in (Figure 1A). At this point, there are two possible options:

- If the cutoff threshold of the initial element distribution (Figure 1A) belongs to the elements above the cutoff (90th percentile) of the bootstrap distribution, then it is admitted as a cutoff threshold for the selection of the significant terms. Since elements comprising ontological terms are sorted in ascending order of significance, only elements greater than the element that corresponds to the cutoff percentage are accepted from the initial distribution.
- If the cutoff threshold of the initial element distribution is below the cutoff of the bootstrap distribution then it cannot be accepted as a cutoff threshold for the significant ontological terms. In that case, the element just over the cutoff threshold of the bootstrap element distribution is taken as the desired threshold instead.

In this way, the terms selected (Figure 1B) are immunized against the bias infiltrating statistical enrichment analyses, producing technically very high statistical scores due to the finite nature of the data population. Besides their high statistical score, the terms gleaned contain a substantial number of biological entities (biomolecules-genes, proteins, etc.) thus incurring a biological prioritization in the selection of the terms, amenable to a Systems Biology context. The output derived, represents a detailed ranked list of significant terms and sets a starting point for further functional analysis.

A mathematically formalized algorithmic description of the aforementioned process for the case of GO terms follows:

```
// Step 0: Import Data
// Step 1: Create the ED
∀GOT_i in DE \{ i=1...|GOT in DE| \}
   a = F(GOT_i in DE)  
   b = F(GOT_i in AL)  
   ED_i = a/b  
   Assign p – value score p_i in ED_i

// Step 2: Sort ED and create the UED
∀z ∈ ED
   f_z = F(z in ED)  
   p_z = p(z)  
Sort ED in descending order by:
   1. f_z  
   2. If f_m = f_n for m,n ∈ {1...|ED|}, m ≠ n
      Sort z_m,z_n by p_z  
End  
UED = Unique(ED)

// Step 3: Define the EID function
Define EID: ED → ℕ such that:
EID(k) = Rank(k) in UED

// Step 4: Create homomorphism between GOTs and EID
Define h: GOT in DE → (ℕ,ℕ) such that:
h(GOT_i) = (i,ID_i) \{ i=1...|GOT in DE| \}

// Step 5: Create the BED, BUED and PED, and find SGOTs
Let A = EID(xth percentile of ED)
For q = 1 to b
   BED_q = Bootstrap(ED)  
   Repeat Step 2 for BED_q and find BUED_q
   PED_q = EID(xth percentile of BED_q)
End
Let B = xth percentile of PED
If A > B
   SGOT = \bigcup_i h^{-1}(i,ID_i > A)
Else
   SGOT = \bigcup_i h^{-1}(i,ID_i > B)
End
```
Chatziioannou and Moulos

Using StRAnGER for ontological analysis

Chao, Moulos, and others have shown that the significance level of kInSHip AmonG ontoloGIcAl termS

Another important, unique feature of StRAnGER, has to do with its capability to assess the level of kinship among the distribution of significant ontological terms, compared to a similar distribution of terms that would be randomly picked by mere chance. This functionality enables a first, overall validation of the results of the analytical procedure at the computational layer, prior to the biological one. In order to examine the significance of the extent of kinship for the resulting population of enriched ontological terms, StRAnGER uses the following, graph-theoretic, pairwise, absolute distance metric (Moulos et al., 2009), of the relevant ontological tree (i.e., for the GO tree, Cellular Component-CP, Molecular Function-MF, or Biological Process-BP):

\[ D_{ij} = \frac{2}{n_i (n_i - 1)} \sum_{p=1}^{n_i} \sum_{q=1}^{n_i} d_{GOT,GOT}^{UAG}, \]

where \( n_i \) denotes the number of ontological terms in either category (for the GO tree CP, MF, or BP) and denotes the absolute distance between nodes \( i \) and \( j \) in the tree undirected acyclic graph (UAG). \( D_{ij} \) represents the average of all pairwise absolute distances among significant ontological terms, UAG calculated for all ontological categories. The statistical validity of \( D_{ij} \) is assessed by resampling the same number of ontological terms from each corresponding ontological category (here GO categories) from the respective reference GO sets, resulting in distinct background bootstrap distributions of \( D_{ij} \)'s for each ontological category. Statistical significance is assessed using the bootstrap \( p \)-value:

\[ \hat{p} = \frac{\#\{D_{ij}^{\text{background}} \leq D_{ij}^{\text{significant}}\}}{m}, \]

where \( m \) is the number of resampling iterations. \( \hat{p} \), represents a modification of the estimation of the Achieved Significance Level for bootstrap (Efron and Tibshirani, 1993).

OUTPUT

StRAnGER supports different output options. The text output of the application is a structured text file, in tab-delimited format that contains the enriched ontological terms resulting from StRAnGER, accompanied by their biological annotations and statistics, such as the statistical test’s \( p \)-value (raw and corrected) and the enrichment ratio. Under each term, the names of the biological entities related to this term are placed, which comprise the significant list (genes, proteins, other molecules), together with further useful information like their unique accessions, names, descriptions and individual \( p \)-values (if provided by the user). In addition, additional information on each entity (e.g., expression values, additional identifiers), can be appended to the output file, if included in the input file with the significant entities, should the user wished so. This information is also given as a webpage (html file), where each GO term is linked with the AmiGO browsing tool for GO (Carbon et al., 2009), each resulting KEGG pathway is linked to the corresponding fully annotated entry (and map) in the KEGG pathways database and each gene is hyperlinked with GeneCards database (Safran et al., 2010), in order to provide additional, versatile, available, annotated information about each entity (i.e., gene). The design of the text output is user friendly, facilitating the interpretation from the side of the expert. The text output can be further configured to comply with

![Figure 1](https://example.com/f1.png)

**Figure 1** (A) Frequency of sorted elements, according to StRAnGER algorithm: the figure presents the number of observations for each element. The dash-dotted line depicts the cutoff element by just applying a threshold on the statistical \( p \)-value, while the dashed line depicts the corrected threshold based on the application of bootstrap. (B) The left bar labeled "Before" depicts the ratio of enriched terms to the number of all terms that the significant genes are annotated to. This ratio is derived after the application of solely an enrichment score \( p \)-value cutoff. The finite nature of the enrichment statistical tests renders the analysis extremely sensitive to false positives, and thus tremendously error prone, due to bias infiltration, as many terms can obtain technically a high statistical score. The right bar labeled "After" depicts the same ratio after the application of StRAnGER algorithm. The graphs are based on the data presented in Moulos et al. (2009).
the following three subtypes: (i) a file containing all information (ontological terms, entities related, statistics), (ii) a file containing only the ontological terms with their statistics ($p$-value, enrichment score), and (iii) a file that contains only the names of the ontological terms. The latter is provided for compatibility or further processing with other applications (e.g., pathway mapping based only on GO terms).

**Visualization**

StRAnGER supports advanced visualization capabilities of its results in both cases, namely GO or KEGG pathway terms analysis. Regarding visualization of GO analysis, a tree-like illustration is adopted depicting the ancestor/descendant relationships among the enriched GO terms, so that correlation among the significant ones can be visualized and conceptualized. The depth of the hierarchical relationships is determined by the user at the beginning. Each significantly enriched GO term is colored. By accentuating (lower $p$-values) or dimming (higher $p$-values) the node color, statistical significance is demonstrated. In this way, the user can easily capture overall the most significant GO terms, at a glance. In addition, the application supports batch mode, appropriate for performing integrative analysis of various datasets together, where the users may upload multiple StRAnGER outputs. In this way a combined graph output is created, able to visualize possible relationships among several experimental outcomes at the functional level. The GO terms from each experiment are colored differently, whereas also terms that are common among different experimental results may be colored elsewise. In this way, the researcher can easily locate biological functions shared in multiple experiments and set points of further investigation. Regarding the visualization of the results of KEGG Pathways, StRAnGER interfaces with web services originally developed by KEGG administration that implement advanced visualization functionalities of KEGG pathway maps, in order to paint selected pathway maps, resulting as significant from its execution, with the related molecular entities (genes, or possibly proteins) that are considered significantly differentially expressed and are present in these pathway illustrations.

**Implementation**

The core StRAnGER routine is programmed in Perl. The GO directed acyclic graph (DAG) is retrieved and handled using tools from MATLAB's 7.4 (R2007a) Bioinformatics Toolbox and converted to an UAG by transforming its adjacency matrix, in order to measure absolute distances between DAG nodes. The visualization of GO terms relationships is achieved by a combination of MATLAB routines and the graphViz library\(^1\). The whole application is built upon a Linux server, running Apache, MySQL, and PHP platforms. At present, and in order to promote the concept of pervasive, distributed computing, a Java programmed StRAnGER web service is currently under development (WSDL representation), to provide availability and access to its functionality, by being incorporated in third party analysis workflows, through the use of appropriate workflow managers like the Taverna Workbench (Hull et al., 2006).

\(^1\)http://www.graphviz.org/

**Results and Discussion**

StRAnGER is a web-based application\(^1\), which performs high-throughput genomic experiment functional analysis, for multiple types of data (i.e., microarray or next generation sequencing experimental datasets). For this scope, it exploits two organizational schemas for standardized biological description: that of GO and that of KEGG pathway database. StRAnGER is unique, regarding its capability of performing either GO analysis or molecular pathway analysis, exploiting the top-rated internationally KEGG pathways collection. In both cases, it performs enrichment analysis, through various statistical tests, coupled with resampling techniques to ensure robustness of the results. StRAnGER supports intelligible graphical representation of the results (Figure 2), whether the enrichment analysis concerns GO or KEGG pathway terms. In this way, the researcher can project and correlate results from different experiments, and possibly extract biologically useful conclusions. Its web-based nature renders StRAnGER platform-independent. Its design is absolutely transparent, enabling easy accessibility to all its features, even for computer beginners. Additionally, its simple, user-friendly interface (Figure 3) empowers the user to complete several analysis rounds fast and accurately.

StRAnGER is fully operative\(^1\). StRAnGER has been successfully used, routinely, for analysis of several microarray datasets, exploiting either GO or KEGG Pathways, among which, datasets concerning human PC-3 prostate cancer cell survival (Tenta et al., 2007) and the effects of mastic oil on mouse lung cancer cells (Moulos et al., 2009). In order to demonstrate the usefulness of StRAnGER in enabling prioritized functional analysis, by deriving a sorted list of GO terms or KEGG pathways, the results of the analysis of the effects of mastic oil on mouse lung cancer cells are presented in brief. Mastic oil, a natural extract from *Pistacia lentiscus* variation *chia* is a blend of bioactive terpenes with identified therapeutic properties (Koutsoudaki et al., 2005). It has been shown to induce anti-tumor activities by possibly inhibiting cancer cell proliferation, survival, angiogenesis, and inflammatory response (Magkouta et al., 2009). In Moulos et al. (2009), StRAnGER was used to infer enriched GO terms, in a designed time-course DNA microarray study, performed to reveal possible anti-tumor effects exerted by mastic oil treatment on mouse Lewis lung carcinoma (LLC) cells. The significantly enriched GO terms according to StRAnGER, are presented in Table 1. The correlation of the enriched GO terms, to specific biological processes, was assessed based on the resampling distance-based algorithm described in Section “Materials and Methods” ($p_t = 0$ and $p_r = 0.001$), supporting the validity of those terms.

From those top performing GO terms according to StRAnGER, emphasis was given to four of them, GO:0008285 (negative regulation of cell proliferation), GO:0007049 (cell cycle), GO:0036917 (induction of apoptosis) and GO:0043123 (positive regulation of 1-XB kinase/NF-XB cascade). Those terms were selected on the basis of including genes with high expression rates but also for their established role in cancer progression supported by previous experimental evidence (Magkouta et al., 2009). The phosphatase and tensin homolog deleted on chromosome ten (Pten) gene was found to be involved in three of the selected GO categories while there is evidence supporting a functional cross-talk between PTEN and NF-XB signaling (Vasudevan et al., 2004). Pten along with three
other genes, E2f7, Nod1, and Hmox1, were validated by RT-PCR analysis in all time points of the experiment, presenting overall good correlation with the microarray analysis profile of differential expression.

Besides GO analysis, the algorithm of StRAnGER was applied in terms of the KEGG Ontology (KO), which emphasize in a detailed description of various aspects of cellular metabolism (KEGG Pathways) across different species. Both approaches are complementary since GO-analysis retrieves functional information concerning cellular physiology in a broader molecular context, while KEGG pathway analysis maps the alterations observed at the -omic experiments, at detailed biochemical cellular reaction networks. Coupled to the interesting cancer-related processes, which were highlighted by GO analysis, and imply perturbation of cell cycle (“negative regulation of cell proliferation,” “DNA replication initiation,” “cell cycle,” “cell division”), KEGG analysis (Table 2) corroborates GO-analysis, regarding glutathione and fatty acid metabolism, DNA-related procedures (purine and pyrimidine metabolism), whereas it gives also emphasis...
to xenobiotic metabolism due to the combined alterations observed in these molecular pathways. Overall, it can be said that the physiological information captured through both analyses, epitomized in the functional description of these terms, reflects a consistent, combined overview of the cancerous physiology, revealing however crucial functional parameters that may aid the expert in gaining insight into the intricacies of its manifestation. Glutathione metabolism for instance is tightly related to the manifestation of antioxidant action and could thus provide missing links for the explanation of the apparent anti-tumorigenic activity of mastic oil. On the other hand, as mastic oil is a plant derived mixture of terpenes, it induces a systemic cellular response, demonstrated by the activation of xenobiotic metabolic pathways. Albeit these findings represent suggestions for further investigation, it is obvious that StRAnGER facilitates the design of targeted experiments, for validation of new hypotheses, as was the case in (Moulos et al., 2009). Yet, the ranked list of terms, accompanied by the individual significant molecular players linked to these terms, and useful information about their expression, constitutes a case-specific valuable pool of scalable, biological information that can be integrated at a controlled pace.

In order to evaluate the performance of StRAnGER compared to other state of the art GO analysis tools, a simple comparison of the results of StRAnGER with those of, GOEAST (Zheng and Wang, 2008) and GOstat (Beissbarth and Speed, 2004) was performed. The data of Moulos et al. (2009) were used for statistical GO analysis through GOEAST and GOstat, applying two different statistical configurations: (i) no multiple-testing correction (ii) Benjamini–Yekutieli correction of FDR under-dependencies. The number of the resulting GO terms for a \( p \)-value cutoff of 0.05 was measured and the results are presented in Figure 4. The default StRAnGER algorithm, with a \( p \)-value cutoff of 0.05 and a 90th percentile cutoff to its element distribution, results in 63 GO terms while GOEAST...
### Table 1 | A list of significantly enriched GO terms derived from the data in (Moulos et al., 2009) after the application of StRAnGER.

| GO term                  | Definition                                                | Set | p-value   | Enrichment |
|--------------------------|-----------------------------------------------------------|-----|-----------|------------|
| GO:0005634 Nucleus       | C                                                         | C   | 1.55E–10  | 194/6236   |
| GO:0016740 Transferase activity | F                                                      | F   | 6.62E–08  | 88/2383    |
| GO:0000166 Nucleotide binding | F                                                     | F   | 8.81E–08  | 99/2891    |
| GO:0016787 Hydrolase activity | F                                                     | F   | 1.16E–06  | 78/2246    |
| GO:0019900 Kinase binding | F                                                        | F   | 1.34E–06  | 4/11       |
| GO:0004402 Histone acetyltransferase activity | F                                                     | F   | 1.74E–06  | 6/30       |
| GO:0003676 Nucleic acid binding | F                                                     | F   | 1.79E–06  | 72/2045    |
| GO:0003723 RNA binding    | F                                                        | F   | 1.89E–06  | 42/978     |
| GO:0046872 Metal ion binding | F                                                      | F   | 2.92E–06  | 109/3528   |
| GO:0005622 Intracellular  | C                                                        | C   | 3.59E–06  | 88/2704    |
| GO:0005515 Protein binding | F                                                      | F   | 5.37E–06  | 198/7381   |
| GO:0003436 Glutathione transferase activity | F                                                     | F   | 6.41E–06  | 6/36       |
| GO:0030509 BMP signaling pathway | P                                                     | P   | 8.16E–06  | 4/15       |
| GO:0003824 Catalytic activity | F                                                     | F   | 8.86E–06  | 35/605     |
| GO:0043433 Negative regulation of transcription factor activity | P                                                     | P   | 1.06E–05  | 3/8        |
| GO:0005524 ATP binding    | F                                                        | F   | 1.07E–05  | 77/2348    |
| GO:0016491 Oxidoreductase activity | F                                                     | F   | 1.18E–05  | 41/1019    |
| GO:0006270 DNA replication initiation | P                                                     | P   | 2.95E–05  | 4/19       |
| GO:0008415 Acyltransferase activity | F                                                     | F   | 5.38E–05  | 14/224     |
| GO:0008270 Zinc ion binding | F                                                        | F   | 5.47E–05  | 93/3114    |
| GO:0000287 Magnesium ion binding | F                                                     | F   | 6.96E–05  | 25/654     |
| GO:0005739 Mitochondrion  | C                                                        | C   | 0.00011592| 45/1275    |
| GO:0016874 Ligase activity | F                                                        | F   | 0.00015613| 21/454     |
| GO:0003954 NADH dehydrogenase activity | F                                                     | F   | 0.000236  | 5/44       |
| GO:0008285 Negative regulation of cell proliferation | P                                                     | P   | 0.00026588| 10/152     |
| GO:0030855 Epithelial cell differentiation | P                                                     | P   | 0.00030166| 4/30       |
| GO:0008137 NADH dehydrogenase (ubiquinone) activity | F                                                     | F   | 0.00030269| 5/46       |
| GO:0030529 Ribonucleoprotein complex | C                                                     | C   | 0.00034927| 20/449     |
| GO:0006412 Translation    | P                                                        | P   | 0.00037692| 21/484     |
| GO:0007049 Cell cycle     | P                                                        | P   | 0.00045196| 27/694     |
| GO:0005762 Mitochondrial large ribosomal subunit | C                                                     | C   | 0.00047806| 4/33       |
| GO:0042157 Lipoprotein metabolic process | P                                                     | P   | 0.00049001| 3/19       |
| GO:0006464 Protein modification process | P                                                     | P   | 0.00052038| 14/276     |
| GO:0008152 Metabolic process | P                                                     | P   | 0.00054429| 32/681     |
| GO:0008134 Transcription factor binding | F                                                     | F   | 0.000618  | 9/142      |
| GO:0016301 Kinase activity | F                                                        | F   | 0.00062479| 45/1378    |
| GO:0003735 Structural constituent of ribosome | F                                                     | F   | 0.00069785| 14/284     |
| GO:0006869 Lipid transport | P                                                        | P   | 0.00071252| 7/96       |
| GO:0006260 DNA replication | P                                                        | P   | 0.00071381| 11/198     |
| GO:0006749 Glutathione metabolic process | P                                                     | P   | 0.0010504 | 3/23       |
| GO:0006631 Fatty acid metabolic process | P                                                     | P   | 0.0011307 | 7/103      |
| GO:0008092 Cytoskeletal protein binding | F                                                     | F   | 0.001226  | 6/81       |
| GO:0006470 Protein amino acid dephosphorylation | P                                                     | P   | 0.0012779 | 12/241     |
| GO:00045177 Apical part of cell | C                                                     | C   | 0.0016441 | 4/43       |
| GO:0006446 Regulation of translational initiation | P                                                     | P   | 0.0016908 | 3/26       |
| GO:0006350 Transcription | P                                                        | P   | 0.0018383 | 64/2230    |
| GO:0005737 Cytoplasm     | C                                                        | C   | 0.0018487 | 66/2314    |
| GO:0043123 Positive regulation of I kappaB kinase/NF kappaB cascade | P                                                     | P   | 0.0019357 | 5/65       |
| GO:0004721 Phosphoprotein phosphatase activity | F                                                     | F   | 0.0021059 | 12/255     |
| GO:0005840 Ribosome      | C                                                        | C   | 0.0021789 | 12/256     |
| GO:0001814 mRNA catabolic process, nonsense-mediated decay | P                                                     | P   | 0.0022434 | 3/28       |
| GO:0006917 Induction of apoptosis | P                                                     | P   | 0.0022511 | 8/141      |
Table 1 | Continued

| GO term     | Definition                                            | Set | p-value   | Enrichment |
|-------------|-------------------------------------------------------|-----|-----------|------------|
| GO:0003677  | DNA binding                                          | F   | 0.0022689 | 82/3014    |
| GO:0051301  | Cell division                                        | P   | 0.0023318 | 14/321     |
| GO:0045944  | Positive regulation of transcription from RNA polymerase II promoter | P   | 0.0024536 | 11/229     |
| GO:0005783  | Endoplasmic reticulum                                | C   | 0.0025159 | 34/1045    |
| GO:0030528  | Transcription regulator activity                      | F   | 0.0026217 | 14/325     |
| GO:0009117  | Nucleotide metabolic process                         | P   | 0.0029513 | 4/49       |
| GO:0006118  | Electron transport                                    | P   | 0.0031794 | 23/645     |
| GO:003713   | Transcription coactivator activity                    | F   | 0.0032915 | 7/122      |
| GO:0004842  | Ubiquitin–protein ligase activity                     | F   | 0.0034684 | 11/239     |
| GO:00058802 | Trans-Golgi network                                   | C   | 0.0034872 | 5/73       |
| GO:0016481  | Negative regulation of transcription                  | P   | 0.0035622 | 8/151      |

The list of 925 significantly modulated genes from Moulos et al. (2009) was submitted to STRAnGER for GO-based meta-analysis, elucidating over-represented, and significantly enriched GO terms. p-value represents the hypergeometric test p-value score for each GO term and enrichment represents the ratio of the number of times a GO term occurs in the significant gene list to the number of times this GO term exists in the list of the entire microarray or the respective background.

Table 2 | A list of significantly enriched KEGG pathways derived from the data in Moulos et al. (2009) after the application of STReNGER.

| KEGG ID | KEGG pathway                                   | Class                                        | p-value   | Enrichment |
|---------|------------------------------------------------|----------------------------------------------|-----------|------------|
| 00603   | Glycosphingolipid biosynthesis – globo series | Metabolism; glycan biosynthesis and metabolism | 2.01E–11  | 1/1        |
| 00670   | One carbon pool by folate                      | Metabolism; metabolism of cofactors and vitamins | 1.32E–05  | 5/27       |
| 00480   | Glutathione metabolism                         | Metabolism; metabolism of other amino acids  | 2.26E–05  | 7/69       |
| 00920   | Sulfur metabolism                              | Metabolism; energy metabolism                | 3.07E–05  | 3/10       |
| 00450   | Selenoamino acid metabolism                    | Metabolism; metabolism of other amino acids  | 3.17E–05  | 2/4        |
| 00230   | Purine metabolism                              | Metabolism; nucleotide metabolism            | 3.59E–05  | 12/167     |
| 00240   | Pyrimidine metabolism                          | Metabolism; nucleotide metabolism            | 8.76E–05  | 5/57       |
| 04130   | SNARE interactions in vesicular transport       | Genetic information processing; folding, sorting and degradation | 0.000294  | 6/64       |
| 00780   | Biotin metabolism                              | Metabolism; metabolism of cofactors and vitamins | 0.000402  | 1/2        |
| 00450   | Selenoamino acid metabolism                    | Metabolism; metabolism of other amino acids  | 0.000429  | 5/49       |
| 00071   | Fatty acid metabolism                           | Metabolism; lipid metabolism                 | 0.000468  | 6/69       |
| 00240   | Pyrimidine metabolism                           | Metabolism; nucleotide metabolism            | 0.000535  | 8/115      |
| 00624   | 1- and 2-Methylnaphthalene degradation          | Metabolism; xenobiotics biodegradation and metabolism | 0.000822  | 4/37       |
| 00362   | Benzoate degradation via hydroxylation          | Metabolism; xenobiotics biodegradation and metabolism | 0.000868  | 2/10       |
| 04720   | Long-term potentiation                          | --                                           | 0.000869  | 8/123      |
| 00980   | Metabolism of xenobiotics by cytochrome P450    | Metabolism; xenobiotics biodegradation and metabolism | 0.000933  | 7/100      |
| 00643   | Styrene degradation                             | Metabolism; xenobiotics biodegradation and metabolism | 0.001189  | 1/3        |

The list of 925 significantly modulated genes from Soong (2004) was submitted to STRAnGER to derive enriched KEGG pathways based on the algorithms described in the text. p-value represents the hypergeometric test p-value score for each KEGG pathway and enrichment represents the ratio of the number of times a KEGG pathway occurs in the significant gene list to the number of times this KEGG term exists in the list of the entire microarray or the respective background.

results in 721 and 133 terms without and with multiple-testing correction respectively, and GOStat results in 359 and 37 terms without and with multiple-testing correction respectively. The results of all executions are provided in Supplementary file 1 online.

A comparison of the results of STRAnGER with both lists derived from GOEAST reveals that in both cases, STRAnGER returns fewer (Figure A1A in Appendix) and more specific in terms of biological function terms, with the prioritization algorithm being able to filter out very generic functions such as “cell part” or “binding.” It is to be noted that even after multiple-testing correction in GOEAST, such terms remain, with some of them representing root nodes in the GO hierarchical tree, e.g., “cell.” Regarding the results of GOstat, it should be stressed that in the case of the uncorrected GO terms list, a plethora of GO terms with enrichment 1/1 are present, which constitute “leaves” of the GO tree and are totally uninformative in the pathway context, (see Supplementary Material), while in the list derived after application of the multiple-testing correction, mainly generic nodes such as “membrane-bounded organelle” are qualified, even if the total number of returned GO terms is lower than the number of terms returned by STRAnGER. This happens because the application of the multiple-testing correction is applied solely in the distribution of p-values, thus applying a stricter statistical threshold. As the classical enrichment tests favor terms enumerating many members, this has a clear impact in the resulting list of terms.
Our presumption, which serves as a performance criterion, is that GO terms representing very specific biological functions, identical in practice with those of one or two genes, lying very low in the GO hierarchy (very poor information content in the pathway context), should be filtered from an output list of GO terms, despite the fact they present an extremely high enrichment. To assess the ability of the three packages to filter these hierarchically low terms, their observations were counted in the outcome lists. The results are shown in Figure 4, both as simple counts (A), as well as normalized to the total number of terms for each outcome (B). It should be noted that even though this scaling step was not entirely necessary, as the algorithm of each program is independent and responsible for the noise included in each outcome, it was performed in order to avoid possible bias in the scoring scheme, caused by large differences in the numbers of enriched GO terms, derived by each application. Moreover, in order to standardize the evaluation of the three solutions, trivial terms presenting a very low gene content (i.e., 1, 2, 3 genes), and at the same time yield a very high enrichment score ($p$-value < 0.001) were filtered out. It should be noted that StrAnGER, by its inner logic, is always capable of eliminating these terms, so these terms were excluded from the comparative analysis, in order to neutralize the evaluation and avoid favoring its results pro StrAnGER. However, still the curve corresponding to StrAnGER (solid line with triangle markers) lies constantly below...
the curves of GOEAST (dash-dotted line with square markers) and GOstat (dashed line with circle markers), stressing out the ability of the prioritization algorithm to filter out hierarchically low GO terms of limited biological significance in a pathway context, without the use of multiple-testing correction. Similar results are acquired for GO terms connected to generic functions (count in background list >2000), where StRAnGER filters out many of them (Figure A1B in Appendix).

Overall, the uniqueness of StRAnGER in forming a ranked distribution of terms, according to their enrichment scores and then applying resampling techniques, mitigates the bias either in favor of very high enrichment scores, or of very general terms, with high gene membership. Overall, it seems that StRAnGER is able to capture a significant portion of crucial biological functions encountered in the dataset, while at the same time, through prioritization, renders the inflow of information, for integration purposes from the resulting list, manageable by the expert.

StRAnGER aspires to provide researchers with rational targets for further research by exploiting either the GO or the KEGG pathway database. However, the methodology is generic enough, to efficiently accommodate in the near future, other biological ontologies and support analysis, by using their annotations. Despite the wealth of tools for ontological analysis, to the best of our knowledge there is no other approach, combining the application of biological prioritization on the terms, based on the types of enrichments observed in the population of terms, with the subsequent, generic use of strict statistical criteria together with bootstrapping, as a means to ensure robustness of the results. StRAnGER provides correction by applying a stringent distribution-free criterion through the use of bootstrapping than just by using the raw statistical test alone. The use of resampling enables the robust automated derivation of the list of significant processes At the same time, p-value correction based on widely used FWER methods (Dudoit et al., 2002) is also provided as additional information in the resulting output list. The resulting list of GO or KEGG Pathway terms and their clear mapping to distinct cellular procedures, introduce a pathway perspective as regards the interpretation of these experiments, thus refactoring the analysis of lists of significant genes, to empower a systems level functional interpretation.

ACKNOWLEDGMENTS

The authors would like to thank Efthimios Ladokukakis for essential contribution in the development of the generic tree visualization and CERVis tools using the graphviz libraries and also Ioannis Kanaris and Ioannis Valavanis for generous help in the development of the PHP front-end for the StRAnGER application.

SUPPLEMENTARY MATERIAL

The Supplementary material for this article can be found online at http://www.frontiersin.org/systems_biology/10.3389/fnins.2011.00008/

REFERENCES

Alexa, A., Rahnenfurther, L., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25, 23–29.

Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. N. (2008). Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24, 1650–1651.

Beissbarth, T., and Speed, T. P. (2004). GOSTat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 20, 1464–1465.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Boyle, E. A., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). GO::TermFinder—a open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics 20, 3710–3715.

Carbon, S., Ireland, A., Mungall, C. J., Shi, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics 25, 288–289.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R.A. (2003). DAVID database for annotation, visualization, and integrated discovery. Genome Biol. 4, P3.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sin. 12, 111–140.

Efron, B., and Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.

Flucke, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Krokociński, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidt, P., Overduin, B., Parker, A., Pritchard, B., Příličný, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vieira, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, M., Herrero, J., Hubbard, T. J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008). Ensembl 2008. Nucleic Acids Res. 36, D707–D714.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitri, G., Smith, C., Smyth, G., Tierney, L., Yang, Y. J., and Zhang, J. (2004). Bioconductor: open source development for computational biology and bioinformatics. Genome Biol. 5, R80.

Goeman, J. J., and Mansmann, U. (2008). Multiple testing performed on the directed acyclic graph of gene ontology. Bioinformatics 24, 537–544.

Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics 23, 3024–3031.

Hosack, D. A., Dennis, G. Jr., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. 4, R70.

Huang, W.-D., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tool: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Poccok, M. R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. Nucleic Acids Res. 34, W729–W732.

Kanchisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–D360.

Koutsoudaki, C., Kesk, M., and Rodger, A. (2005). Chemical composition and antibacterial activity of the essential oil and the gum of Pistacia lentiscus Var. chia. J. Agric. Food Chem. 53, 7681–7685.

Lee, H. K., Braynen, W., Keshav, K., and Pavlidis, P. (2005). Ermine: tool for functional analysis of gene expression data sets. BMC Bioinformatics 6, 269. doi:10.1186/1471-2105-6-269

Magkouta, S., Stathopoulou, G. T., Psallidas, I., Papapetropoulos, A., Kolissi, F. N., Rousos, C., and Loutrari, H. (2009). Protective effects of mastic oil from Pistacia lentiscus variation
Chatziioannou and Moulos Using StRAnGER for ontological analysis

chia against experimental growth of Lewis lung carcinoma. *Nutr. Cancer* 61, 640–648. 

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biol.* 5, R101.

Maynard, S. E., Min, J. Y., Merchan, J., Lim, K. H., Li, J., Mondal, S., Libermann, T. A., Morgan, J. P., Sellke, F. W., Stillman, I. E., Epstein, F. H., Sukhatme, V. P., and Karumanchi, S. A. (2003). Excess placental soluble fms-like tyrosine kinase 1 (sFlt1) may contribute to endothelial dysfunction, hypertension, and proteinuria in preeclampsia. *J. Clin. Invest.* 111, 649–658.

Moulos, P., Papadodima, O., Chatziioannou, A., Loutrari, H., Roussos, C., and Kolisis, F.N. (2009). A transcriptomic computational analysis of mastic oil-treated Lewis lung carcinomas reveals molecular mechanisms targeting tumor cell growth and survival. *BMC Med. Genomics* 2, 68. doi: 10.1186/1755-8794-2-68

Osier, M. V., Zhao, H., and Cheung, K. H. (2004). Handling multiple testing while interpreting microarrays with the gene ontology database. *BMC Bioinformatics* 5, 124. doi: 10.1186/1471-2105-5-124

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010, baq020.

Soong, T. T. (2004). *Fundamentals of Probability and Statistics for Engineers*. Hoboken, NJ: John Wiley & Sons.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Tenta, R., Katopodis, H., Chatziioannou, A., Pilalis, K., Calvo, E., Luu-The, V., Labrie, F., Kolisis, F., and Koutsilieris, M. (2007). Microarray analysis of survival pathways in human PC-3 prostate cancer cells. *Cancer Genomics Proteomics* 4, 309–318.

Zheng, Q., and Wang, X. J. (2008). GOEAST: a web-based software toolkit for gene ontology enrichment analysis. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Tarentino, R., Katopodis, H., Chatziioannou, A., Loutrari, H., Calvo, E., Luu-The, V., Labrie, F., Kolisis, F., and Koutsilieris, M. (2007). Microarray analysis of survival pathways in human PC-3 prostate cancer cells. *Cancer Genomics Proteomics* 4, 309–318.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Tarca, A. L., Romero, R., and Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *Am. J. Obstet. Gynecol.* 195, 373–388.

Tenta, R., Katopodis, H., Chatziioannou, A., Pilalis, K., Calvo, E., Luu-The, V., Labrie, F., Kolisis, F., and Koutsilieris, M. (2007). Microarray analysis of survival pathways in human PC-3 prostate cancer cells. *Cancer Genomics Proteomics* 4, 309–318.

Maynard, S. E., Min, J. Y., Merchan, J., Lim, K. H., Li, J., Mondal, S., Libermann, T. A., Morgan, J. P., Sellke, F. W., Stillman, I. E., Epstein, F. H., Sukhatme, V. P., and Karumanchi, S. A. (2003). Excess placental soluble fms-like tyrosine kinase 1 (sFlt1) may contribute to endothelial dysfunction, hypertension, and proteinuria in preeclampsia. *J. Clin. Invest.* 111, 649–658.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 July 2010; accepted: 11 January 2011; published online: 26 January 2011.

Copyright © 2011 Chatziioannou and Moulos. This is an open-access article subject to an exclusive license agreement between the authors and Frontiers Media SA, which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors and source are credited.
APPENDIX

**Figure A1** (A). The number of statistically significant GO terms after (light bars) or prior (dark bars) the application of multiple testing correction, as derived by the use of the 3 software packages described in the main text. StRAnGER's main algorithm does not necessarily require the application of a multiple testing procedure, as the bootstrapping applied estimates the GO term distribution cutoffs. (B). Number of the resulting, statistically significant GO terms, representing very general biological functions (i.e. “protein binding”) yielded by the 3 software packages described in the main text. StRAnGER's main algorithm manages to filter out many of them.