Prioritized Functional Analysis of Biological Experiments Using Resampling and Noise Control Methodologies

Eleftherios D. Pilalis and Aristotelis A. Chatziioannou, *Member, IEEE*

Abstract—**StRAnGER is a web application for the automated statistical analysis of annotated experiments, exploiting controlled biological vocabularies, like the Gene Ontology or the KEGG pathways terms. In the first version, StRAnGER featured various gene profiling platforms for functional analysis of genomic datasets, starting from a list of significant genes derived from statistical and empirical thresholds. In the current version, various major improvements have been implemented, namely a new ranking algorithm, the expansion of background distributions with protein annotations, the addition of a mode for batch experiments and a noise-control analysis that evaluates the robustness of the prioritized terms through iterative addition of random genes. Overall, StRAnGER enables a systems level functional interpretation through the utilization of bootstrapping techniques and the detection of distribution-independent term enrichments.**

I. Introduction

VARIOUS high through-put biological experiments yield output data in form of lists of molecular components (genes, proteins etc) that are implicated in the underlying biological processes. These subsets of key molecular components are commonly submitted to statistical methodologies that assess their functional roles based on the enrichment of various annotations, for instance Gene Ontology (GO) terms and pathways (KEGG Pathways identifiers). The Gene Ontology [1] provides such functional annotation adopting a hierarchical schema. In addition, the Kyoto encyclopedia of genes and genomes (KEGG) biological pathway database [2] 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. Thus, these controlled biological vocabularies constitute valuable sources of standardized biological information.

Measuring the comparative gene or protein expression has a critical importance in the functional analysis of biological control mechanisms, phenotyping, and molecular profiling of diseases. In order to highlight statistically significant and biologically relevant molecular players, enabling thus a systemic perspective, emphasis must be given in the involvement of molecular pathways rather than isolated factors, in order to uncover molecular groups that participate in or regulate the same cellular pathway.

In order to perform pathway analysis, the subsets of differentially expressed genes or proteins are usually submitted to software tools, which comprise implementations of statistical tests that estimate the number of successes in a sequence of draws from a finite population without replacement, like the hypergeometric distribution. The statistical tests are thus used to determine the extent of over-representation of a term in the initial list of differential expression, compared to mere chance. As the number of the ontological terms in each dataset may vary from several tens to hundreds or even thousands of terms, the probability for false positives rises, and this requires application of several, possible, multiple correction methodologies [3].

The fundamental presumption of StRAnGER [4], which serves as a performance criterion, is that terms representing very specific biological functions, identical in practice with those of one or two genes, lying very low in the vocabulary hierarchy (very poor information content in the pathway context), should be filtered out from an output list of significant terms, despite the fact they eventually present an extremely high enrichment. To address this problem, StRAnGER detects and prioritizes statistically significant enrichments, instead of only terms, by applying bootstrapping in a distribution of grouped terms by their common enrichments. In this way, the terms selected 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 output terms contain a substantial number of biological entities thus incurring a biological prioritization in the selection of the terms, amenable to a Systems Biology context. In the current version of StRAnGER (available at http://grissom.gr/stranger2), an improved bootstrap ranking algorithm in combination with an iterative process that evaluates the robustness of the enrichments against introduction of noise, strongly enhances the prioritization of systemic biological processes and the overall power of the application in molecular pathway analysis.

Manuscript received July 30, 2013. This work was supported by the "Cooperation" program 09SYN-11-675 (DAMP), O.P. Competitiveness & Entrepreneurship (EPAN II).

E. D. Pilalis is with the Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece (e-mail: epilalis@eie.gr).

A. A. Chatziioannou is with the Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece (corresponding author to provide phone: +302107273751, e-mail: achatzi@eie.gr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

II. STRANGER NEW FEATURES

A. Ranking algorithm

StRAnGER performs enrichment analysis of input terms compared to a reference set, by three statistical tests (hypergeometric test, Fisher's exact test, chi-square test). Subsequently, StRAnGER repartitions and reorders the initial distribution of terms to define a new distribution of "elements". Elements are defined as groups of terms with the same enrichment, pooled together. The application of bootstrap resampling to the distribution of elements provides a corrected measure of the statistical significance of those elements, based on the observation score of particular enrichments.

In the first version of StRAnGER [4], elements with the same frequency were being sorted again in a decreasing order, according their *p*-value enrichment scores and a distribution of the ranked thresholds was being built and compared with the cutoff threshold of the initial element distribution. The final *p*-value cutoff threshold was then decided according to this comparison. If the cutoff threshold of the initial element distribution belonged to the elements above the cutoff (90th percentile) of the bootstrap distribution, then it was admitted as a cutoff threshold, whereas, if the cutoff threshold of the initial element distribution was below the cutoff of the bootstrap distribution, then the element just over the cutoff threshold of the bootstrap element distribution was taken as the desired threshold.

In the current version, a new bootstrap-corrected *p*-value is calculated from the distribution of elements according to the frequency of occurrence of each enrichment score on each position of the elements distribution, after many bootstrapping rounds (default 10000). Then, two separate *p*-value cutoff thresholds are applied: the statistical test *p*-value threshold and the bootstrap-corrected *p*-value of the element above the 90th percentile of the ranked bootstrapping distribution. The elements are derived as statistically significant if they satisfy both of the aforementioned conditions, but they are ranked according to their bootstrapping *p*-value, which is considered as immunized against the bias of the very high scores derived from the statistical tests, due to the finite nature of the data population. The final output of significant terms is spanned by the terms that are mapped to the significant elements.

B. Batch mode

A new mode of execution of the StRAnGER analysis was added, which enables the processing of multiple lists that are commonly derived from batch experiments, e.g. time-course differential gene expression. StRAnGER derives a list of statistically significant terms across different conditions, ranked by their frequency of occurrence. Thus, the most important molecular pathways, which are not specific to a unique time point or condition, are highlighted.

C. Noise mitigation analysis

One of the most important challenges in biological experiments that derive lists of differentially affected molecular components is the control of the inherent noise caused by fluctuations irrelevant to the condition that is being tested. The amounts of nucleic acids, proteins and other molecules very as a function of time, cell cycle, and cell to cell variability [5]. StRAnGER addresses this problem by incorporating a second batch analysis mode, where the robustness of the statistical scores against random noise is assessed. The analysis is performed in multiple rounds by adding random genes to the initially submitted list. The ranks of the terms are stored in each round and each term receives a score which is calculated as the sum of all ranks. The terms are finally ranked according to their scores, from the lowest to the highest. Consequently, terms corresponding to elements that are found sporadically as highly ranked are considered as occasionally prioritized by chance and are filtered out. In contrast, terms that are constantly highly prioritized are robust to random enrichment occurrences and are kept as biologically important.

As a demonstration of this approach, pre-configured analyses are enabled in the StRAnGER web interface, as optional KEGG pathway ids, in order to automatically perform an analysis by submitting lists of genes corresponding to KEGG pathways. An example analysis of the KEGG pathway 00020 (Citrate Cycle) is summarized in Tables I and II. In each round, 300 random genes were added to the initial list of the thirty genes of the Citrate Cycle pathway and five genes effectively belonging to the pathway were randomly selected to be removed. Term GO:0006099 corresponding to Citrate Cycle was robust until round 5 where only 7 genes annotated to this term were left in the input list, in a set of 1210 genes. The final top ten ranked GO terms are shown in table II. Term GO:0006099 was prioritized as overall most important, unambiguously highlighting the citrate cycle pathway. The rest of the terms are absolutely relevant to this pathway, as they describe overlapping or immediately up- or down-stream metabolic functions.

D. Web application

TABLE I
NOISE MITIGATION ANALYSIS OF THE KEGG PATHWAY CITRATE CYCLE

Round	Number of genes	Number of pathway genes	Enrichment of GO:0006099	Rank of GO:0006099
1	30	30	22/26	4
2	325	25	19/26	2
3	620	20	14/26	1
4	915	15	13/26	1
5	1210	10	7/26	1
6	1505	5	3/26	-

GO:0006099 : tricarboxylic acid (citrate) cycle. The analysis was performed in six rounds, adding 300 random genes and removing five random pathway genes, in each round. Hypergeometric p-value cutoff: 0.001. Bootstrap cutoff threshold: p-value of 90th percentile.

TABLE II
RESULTS OF NOISE MITIGATION ANALYSIS

Rank	Frequency	GO id	GO Description	Rounds	Ranks
1	5	GO:0006099	tricarboxylic acid cycle	1 2 3 4 5	4 2 1 1 1
2	3	GO:0005749	mitochondrial respiratory chain complex II	1 2 3	3 1 4
3	3	GO:0006107	oxaloacetate metabolic process	1 2 3	9 13 3
4	3	GO:0010510	regulation of acetyl-CoA biosynthetic process from pyruvate	1 2 3	6 14 10
5	3	GO:0006090	pyruvate metabolic process	1 2 3	8 17 11
6	2	GO:0045254	pyruvate dehydrogenase complex	1 2	2 7
7	2	GO:0006102	isocitrate metabolic process	2 3	4 6
8	2	GO:0006103	2-oxoglutarate metabolic process	2 3	5 7
9	2	GO:0006104	succinyl-CoA metabolic process	1 2	1 12
10	2	GO:0044281	small molecule metabolic process	1 2	5 8

GO:0006099 : Citrate Cycle. The noise mitigation analysis highlighted the robustness of the main metabolic pathway (citrate cycle) through iterative additions of random genes and removal of pathway genes. Relevant terms describing overlapping or immediately up- or down-stream metabolic functions were prioritized as well.

A new web application was implemented using the Web2py framework (available at http://grissom.gr/stranger2). GO Term background distributions were added, comprising Ensemble gene annotations and Uniprot protein annotations for a number of model organisms (human, mouse, rat, thale cress, yeast, E. coli) in order to enable generic executions from data not necessarily derived from microarray platforms, e.g. from next-generation sequencing experiments.

In addition to the web interface, XML and JSON web services were enabled, permitting the integration of StRAnGER algorithms to extended workflows, through any programming language. Consequently, automatic large-scale analyses can be performed in the context of data-intensive, high throughput projects.

III. Conclusion

StRAnGER effectively controls gene/protein expression noise and prioritizes the terms that correspond to molecular pathways enriched in large lists deriving from high throughput experiments. The performance of StRAnGER renders possible the effective pathway analysis in the context of Systems Biology experiments, where the combined functionality of groups of molecules is under investigation, instead of giving emphasis to individual actors. StRAnGER, through the use of distribution-independent bootstrapping correction, derives the statistically important enrichments by avoiding the noise-sensitive, overestimated scores of statistical tests alone, and enables a systems level functional interpretation.

Further development of StRAnGER will comprise the extension of available terms with many kinds of structural and functional annotations (e.g. InterPro protein domains), as the algorithm of StRAnGER permits the abstraction in matter of the utilized vocabularies and the detection of distribution-independent enrichments. Finally, information and entropy-based criteria will be implemented in the decision of the applied thresholds, offering objective measures of the underlying pathway complexity.

References

[1] Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

[2] Kanehisa, M., et al., KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38(Database issue): p. D355-60.

[3] Osier, M.V., H. Zhao, and K.H. Cheung, Handling multiple testing while interpreting microarrays with the Gene Ontology Database. BMC Bioinformatics, 2004. 5: p. 124.

[4] Chatziioannou, A.A. and P. Moulos, Exploiting Statistical Methodologies and Controlled Vocabularies for Prioritized Functional Analysis of Genomic Experiments: the StRAnGER Web Application. Front Neurosci. 2011 5: p. 8.

[5] Sanchez, A., S. Choubey, and J. Kondev, Regulation of noise in gene expression. Annu Rev Biophys, 2013. 42: p. 469-91.