Onto-Tools: new additions and improvements in 2006

Purvesh Khatri1, Calin Voichita1, Khalid Kattan1, Nadeem Ansari1, Avani Khatri1, Constantin Georgescu1, Adi L. Tarca2 and Sorin Draghici1,∗

1Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI, 48202 and 2Perinatology Research Branch-NIH/NICHD, 4 Brush, 3990 John R, Detroit, MI 48201, USA

Received January 31, 2007; Revised April 10, 2007; Accepted April 17, 2007

ABSTRACT

Onto-Tools is a freely available web-accessible software suite, composed of an annotation database and nine complementary data-mining tools. This article describes a new tool, Onto-Express-to-go (OE2GO), as well as some new features implemented in Pathway-Express and Onto-Miner over the past year. Pathway-Express (PE) has been enhanced to identify significantly perturbed pathways in a given condition using the differentially expressed genes in the input. OE2GO is a tool for functional profiling using custom annotations. The development of this tool was aimed at the researchers working with organisms for which annotations are not yet available in the public domain. OE2GO allows researchers to use either annotation data from the Onto-Tools database, or their own custom annotations. By removing the necessity to use any specific database, OE2GO makes the functional profiling available for all organisms, with annotations using any ontology. The Onto-Tools are freely available at http://vortex.cs.wayne.edu/projects.htm.

INTRODUCTION

Together with the ability of generating a large amount of data per experiment, high-throughput technologies also brought the challenge of translating such data into a better understanding of the underlying biological phenomena. First released in 2001, Onto-Tools is a freely available web-accessible software suite that addresses some of these challenges. The Onto-Tools suite includes: (i) Onto-Express—used to translate lists of differentially regulated genes into a better understanding of the underlying biological phenomena (1–5); (ii) Onto-Design—used to select the best set of genes to be included on a custom microarray designed for the study of a given biological phenomenon (2,4); (iii) Onto-Compare—used to analyze the functional bias of various focused commercial microarrays and select the one that is most appropriate for a given biological hypothesis (2,6); (iv) Onto-Translate—used to translate lists of genes from one reference system to another (e.g. from GenBank accession numbers to UniGene cluster IDs to Affymetrix probe IDs, etc.) (2,5,7,8); (v) Onto-Miner—provides a unified access point and an application programming interface (API) allowing queries for various information such as the gene name, official symbol, reference accession number, coded protein, etc. (4); (vi) Promoter-Express—which allows the users to find condition-specific transcription factor binding sites (TFBSs) (7,9) and (vii) nsSNPCounter—which allows analysis of synonymous and non-synonymous codon substitutions in protein coding genes (7). Previous publications have described in detail the motivation, implementation and validation of these tools. The logical workflow between the Onto-Tools applications has also been previously explained (2,4). This article describes two new tools added to the ensemble and discusses other enhancements recently made to the existing tools.

OE2GO

Onto-Express (OE) is a web-based tool in the Onto-Tools suite that performs automated function profiling for a list of differentially expressed genes. However, Onto-Express does not support functional profiling for the organisms that do not have annotations in public domain, or use of custom (i.e. user-defined) ontologies. This limitation is also true for most of the other existing tools for functional profiling (10), which means that researchers working with uncommon organisms and/or new annotations or ontologies may be forced to construct such profiles manually.

Onto-Express-to-go (OE2GO) is a new tool added to the Onto-Tools ensemble to address these issues. OE2GO is built on top of OE to leverage its existing functionality. In OE2GO, the users now have an option to use either the Onto-Tools database as a source of functional annotations or provide their own annotations in a separate file (Figure 1). Currently, OE2GO supports annotation file in the Gene Ontology format. A GO-formatted annotation file has 15 tab-delimited columns that contain a database

*To whom correspondence should be addressed. Tel: +1 313 577 5484; Fax: +1 313 577 6868; Email: sorin@wayne.edu

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
name, a unique ID for an entity being annotated, its
corresponding symbol, one or more references supporting
an annotation, type of evidence, date of annotation, type
of entity (e.g. a gene or a protein), etc. The detailed
description of the format and each column is available at
www.geneontology.org/GO.annotation.shtml. The anno-
tation files for approximately 40 organisms in GO format
are available for download from GO ftp site that can be
directly used with OE2GO (ftp://ftp.geneontology.org/
pub/go/gene-associations/).

As shown in Figure 1, when using custom annotations,
the user must also specify the ontology file in OBO format.
OE2GO assumes that the identifiers in the input file are of the same type as those used in the
annotations file.

The automated functional profiling approach, first pro-
posed by Onto-Express in 2001, has now become the
de facto standard in the second stage analysis of gene
expression data (10). A large number of tools performing
similar ontological analysis are available today. Although
this approach is widely adopted, it considers each
biological process independent of the others, and ignores
dependencies and interactions among them (10).

PATHWAY-EXPRESS

The automated functional profiling approach, first pro-
posed by Onto-Express in 2001, has now become the
de facto standard in the second stage analysis of gene
expression data (10). A large number of tools performing
similar ontological analysis are available today. Although
this approach is widely adopted, it considers each
biological process independent of the others, and ignores
dependencies and interactions among them (10).

At the same time, a number of pathway databases
available in public domain describe how genes interact
with each other in metabolic and signaling pathways
(e.g. KEGG (11), BioCarta (www.biocarta.com),
Reactome (12), etc.). Several tools that allow researchers
to reveal the pathways associated with a given set of
differentially expressed genes already exist (13–23).

Pathway-Express (PE) is a tool in the Onto-Tools
ensemble that is designed to perform a pathway analysis.
When a user submits a list of genes, PE searches the Onto-
Tools database and builds a list of all associated
pathways. The Onto-Tools database currently contains
signaling pathways from KEGG. However, PE can
analyze any collection of pathways described in SBML
(24). PE performs a classical enrichment analysis based on
a hypergeometric distribution in order to identify those
pathways that contain a proportion of differentially
expressed genes that is significantly different from what
is expected just by chance. This analysis produces a set of
P-values that characterize the significance of the pathway
from this statistical perspective (a lower P-value corre-
sponds to a higher significance).

PE also calculates a perturbation factor PF(g) for each
gene on each pathway. This perturbation factor takes into
consideration the (i) normalized fold change of the gene

Figure 1. OE2GO input interface for using custom annotations. When
the user selects ‘my own annotations’, OE2GO allows the user to
specify the name of the annotation file in GO format and the ontology
file in OBO format. When using custom annotations, OE2GO assumes
that the identifiers in the input file are of the same type as those used in
the annotations file.

Figure 2. An example OBO formatted ontology file that contains a
header, a term stanza and a typedef stanza. Each stanza starts with either
[Term] or [Typedef]. The header and the stanzas contain a set of
tag-value pairs in <tag>=<value> format. The complete list of
allowed tags for each type of stanza and the header are described on
GO web site.
and (ii) the number and amount of perturbation of genes upstream of it (i.e. its position on a pathway). The users can use ‘Advanced Options’ button to specify different weights for different types of interactions between genes on the pathways (Figure 3). As shown in Figure 3, PE uses negative weights for ‘inhibition’ and ‘repression’. This gene perturbation factor reflects the relative importance of each differentially expressed gene on the pathway. The impact factor of the entire pathway is calculated using a probabilistic term that takes into consideration the proportion of differentially expressed genes on the pathway and gene perturbation factors of all genes in the pathway. More details about the gene perturbation factors and pathway impact factors, a comparison with the existing methods, and a full discussion of the advantages and disadvantages of these methods are described elsewhere.

When a user submits a list of input IDs, PE converts the list into a list of gene IDs using the Onto-Tools database. The Onto-Tools database integrates a number of public databases including GenBank, dbEST, UniGene, Entrez Gene, RefSeq, KEGG, etc. After creating a list of gene IDs, PE searches the KEGG pathways in the Onto-Tools database for each input gene, and builds a list of pathways containing at least one input gene. Note that the pathways returned by PE depend on the annotations available in KEGG. If a gene is known to be involved in a pathway, but is not annotated as such in KEGG, PE does not return the pathway in its output.

The output of PE is shown in Figure 4. The top left panel displays detailed results for each pathway including: number of input genes and total number of genes on a pathway, probability of obtaining the same number of genes on a given pathway by random chance, impact factor and probability of obtaining the impact factor by random chance for a given pathway, etc. The bar graph in panel A can be sorted in increasing or decreasing order of any column by clicking on the corresponding column header. The bar graph in panel A, pathway details in panel B and input details in panel C are synchronized with each other. For instance, in Figure 4, selecting the apoptosis pathway in panel B, highlights the corresponding bar in red color in panel A and also selects the input genes in panel C (i.e. genes FADD and RELA).

Right-clicking a mouse in the PE output window brings up a context-sensitive popup menu (Figure 4). For instance, the menu displayed by clicking a mouse on a pathway name allows the user to view the corresponding KEGG pathway diagram, download and specify a GML viewer to use with PE, etc. The KEGG pathway diagram that corresponds to the selected pathway can also be viewed by double-clicking the pathway name in either panel A or panel B. PE highlights the input genes in red (up-regulated genes) or blue (down-regulated) in the KEGG diagram (Figure 5). The popup menu also allows to save a pathway in GML format that can be viewed in any GML viewer. This can be done by selecting ‘Save in GML format’ from the popup menu. If a program able to read GML files (e.g. yEd (www.yed.com) or Cytoscape (21)) is already installed on the local machine, the user can specify its location by selecting ‘Set GML viewer command’ from the popup menu. After specifying the GML viewer, a user can select ‘Show pathway graph’ to access GML representation of the pathway from within PE. All tables in the PE output can be saved as a tab-delimited text file by selecting ‘Save this table’ from the popup menu.

Figure 5 also shows the internal representation of the apoptosis pathway (read from the GML file) with the two input genes represented by elliptic nodes. PE’s internal representation also shows how perturbation introduced by the genes FADD and RELA is propagated throughout the pathway in the given condition. The perturbed genes on the apoptosis pathway are shown with colored background, whereas the unperturbed genes are shown with white background, and the direction of the propagation is indicated by an arrow between two genes. In Figure 5, notice that only the area of the pathway downstream of the input genes is perturbed, while the rest of the pathway is unperturbed.

**ENHANCEMENTS**

Over the past year, Onto-Miner (OM) has been reimplemented as a Java tool to make its interface user friendly and consistent with the Onto-Tools suite. The previous HTML input interface of OM is replaced by a new Java interface that is easier to use (Figure 6). Unlike previous version of OM that required the user to
manually download the results from the server, OM now automatically downloads the result file from the Onto-Tools server and saves it in the location specified by the user.

ACKNOWLEDGMENTS

This material is based upon work supported by the following grants: NSF DBI-0234806, CCF-0438970, 1R01HG003491-01A1, 1U01CA117478-01, 1R21CA10740-01, 1R01NS045207-01, 5R21EB000990-03 and 2P30 CA022453-24. Onto-Tools currently runs on equipment provided by Sun Microsystems EDU 7824-02344-U and by NIH(NCRR) 1S10 RR017857-01. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NIH, DOD or any other of the funding agencies. Funding to pay the Open Access publication charges for this article was provided by NSF DBI-0234806.

Conflict of interest statement. None declared.

REFERENCES

Figure 5. The apoptosis pathway as described by KEGG (panel A) and its internal representation, as constructed by Pathway-Express (panel B). PE can save this internal representation in a GML file that can be exported to any other visualization application. In this internal representation, each node corresponds to a gene, and each edge represents an interaction between two genes. In this example, the up-regulated input gene FADD is highlighted with red color and down-regulated gene RELA is highlighted with blue color in both panels. The internal graph also shows how PE propagates the perturbation through the pathway, where the direction of an arrow represents the direction of propagation. Note that in some cases, a gene in KEGG diagram may represent a group of genes. For instance, gene PI3K in the KEGG diagram in fact corresponds to a group of eight nodes in PE’s internal representation: PIK3R5, PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2 and PIK3R3.


