

# Transcriptator: Computational Pipeline to Annotate Transcripts and Assembled Reads from RNA-Seq Data

Kumar Parijat Tripathi<sup>\*,\*\*</sup>, Daniela Evangelista<sup>\*\*</sup>, Raffaele Cassandra,  
and Mario R. Guarracino

Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP),  
High Performance Computing and Networking Institute (ICAR),  
National Research Council of Italy (CNR), Via Pietro Castellino 111, Napoli, Italy  
kumpar@na.icar.cnr.it

**Abstract.** RNA-Seq is a new tool, which utilizes high-throughput sequencing to measure RNA transcript counts at an extraordinary accuracy. It provides quantitative means to explore the transcriptome of an organism of interest. However, interpreting this extremely large data coming out from RNA-Seq into biological knowledge is a problem, and biologist-friendly tools to analyze them are lacking. In our lab, we develop a Transcriptator web application based on a computational Python pipeline with a user-friendly Java interface. This pipeline uses the web services available for BLAST (Basis Local Search Alignment Tool), QuickGO and DAVID (Database for Annotation, Visualization and Integrated Discovery) tools. It offers a report on statistical analysis of functional and gene ontology annotation enrichment. It enables a biologist to identify enriched biological themes, particularly Gene Ontology (GO) terms related to biological process, molecular functions and cellular locations. It clusters the transcripts based on functional annotation and generates a tabular report for functional and gene ontology annotation for every single transcript submitted to our web server. Implementation of QuickGo web-services in our pipeline enable users to carry out GO-Slim analysis. Finally, it generates easy to read tables and interactive charts for better understanding of the data. The pipeline is modular in nature, and provides an opportunity to add new plugins in the future. Web application is freely available at: [www-labgtp.na.icar.cnr.it:8080/Transcriptator](http://www-labgtp.na.icar.cnr.it:8080/Transcriptator).

**Keywords:** RNA-Seq, QuickGO, DAVID, web-services, Python.

## 1 Scientific Background

The advent of new technologies in transcriptome studies such as RNA-Seq changes the face of traditional biological research approaches. Instead of studying one or

---

\* Corresponding author.

\*\* These authors contributed equally to this work.

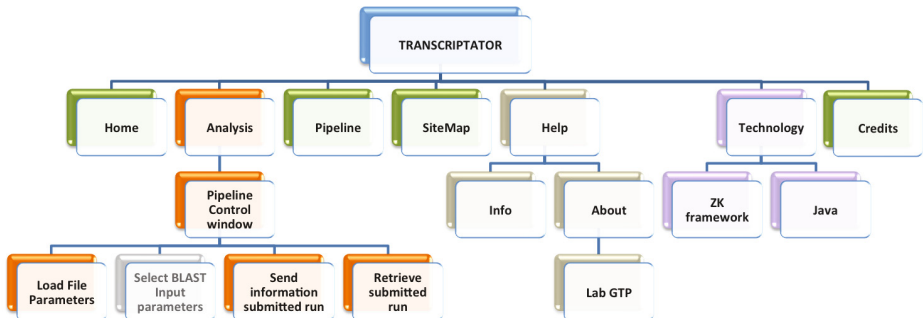
more genes at a time, researchers now simultaneously measure the genome wide changes and regulation of genes under a certain condition. RNA-Seq generates a large amount of biological data in the form of reads. There is a wide array of methodologies to computationally reconstruct the transcript structure and quantify it from raw reads [1]. However, interpreting this extremely large data into biological knowledge is still a challenging and daunting task. A large number of functional annotation pipelines and databases such as DAVID [2], QuickGO [4], ESTExplorer [8], FastAnnotator [5] and other methods [7], were independently developed to address the challenge of functional annotation of the large gene list coming out from RNA-Seq experiments. Both DAVID and QuickGO are very comprehensive databases and can provide putative functional and gene ontological term annotation for a transcript, based on sequence similarity to known genes. These are useful tools for understanding the biological inference of transcriptional response, as well as newly explored sequences. Despite their complex functionalities, both DAVID and QuickGO usually require many manual steps that are often not easy to implement for biologists who are unfamiliar with command line input. Previously, researchers also developed web tools such as FASTAnnotator [5] and ESTExplorer [8]. While ESTExplorer pipeline is specifically designed for EST analysis that includes the cleaning, assembly and clustering and functional annotation of ESTs, FASTAnnotator performs the GO term, enzyme and domain annotations on transcripts. These analyses are not comprehensive as they do not include annotations for pathways such as KEGG, Panther, BioCarta, and they also do not provide any information on protein-protein interactions and other functionalities. They also do not elaborate enrichment analysis for the functional annotation term for the given transcripts data set. The complex plethora of annotation tools and pipelines produces a confusing situation in front of the end users in deciding the most suitable enrichment tool for their analytic skills [3]. There is a need to develop a computational automated pipeline, with a user friendly interface which effectively translates assembled reads coming out from RNA-Seq experiments into biological interpretations such as functional annotation and GO enrichment analysis. We develop a computational pipeline to functional annotate an individual (differentially expressed) transcript, and carry out GO enrichment analysis of expression profiles, under the different treatment condition for organisms, which lacks the referenced genome. In this pipeline, we utilize the web services available for BLAST, QuickGO and DAVID tool for functional and gene ontology annotation and enrichment analysis. Our pipeline carries out automated BLAST run on the Refseq, Swiss-Prot and UniProt-TrEMBL databases to find the most similar genes/proteins for the assembled transcripts. Then, functional and gene ontology annotations are carried out by QuickGo [9] and DAVID web services [6]. The advantage of our pipeline is that it is very easy to use and informative in nature. It produces functional as well as gene ontological annotation for the given transcripts data set. It integrates the results from well established DAVID and QuickGO tools through web services. Our pipeline also provides a plethora of information about enriched pathways such as KEGG, Panther, and BioCarta. The pipeline offers a report on statistical analysis of GO enrichment. It enables a biologist to

identify enriched biological themes, particularly GO terms related to biological process, molecular functions and cellular locations. It also provides information about the SMART, Panther, Prosite, Prodom, PFAM and InterPro domains along with protein interactions such as Mint, Bind for the annotated transcripts. Through our pipeline, we are providing an automated protocol to cluster differentially expressed transcripts based on functional annotation. It is modular in nature, so that it also provides a space for adding up new plugins in the future. All these utilities in our pipeline, deliver a platform for the biologist, to understand the humongous RNA-Seq data in a biological sense and in a straightforward way.

## 2 Materials and Methods

### 2.1 Web Interface

Trascriptator web application is designed using ZK framework (<http://www.zkoss.org/download/zk>) and J2EE (Java 2 Platform Enterprise Edition, [www.oracle.com/technetwork/java/javae](http://www.oracle.com/technetwork/java/javae)) technologies. The modular and distributed J2EE platform is employed to integrate technologies for the exchange of information between different applications, such as XML and Web Services. The implementation of the graphical user interface (GUI) is obtained using ZK framework, Ajax web application open-source, with XUL/XHTML (XML User Interface Language/Extensible HyperText Markup Language) built-in based components. JFASTA library v. 2.1.2 (<http://jfasta.sourceforge.net/>) is implemented in order to handle FASTA format files (.fa). BIOJAVA3-ws module (<http://www.biojava.org/docs/api/org/biojava3/ws>) of BIOJAVA v. 3.0.7 API is used to provide analytical and statistical routines, sequences manipulation -such as BLAST alignment. Lastly, the Jython interpreter v. 2.5.3. (<http://www.jython.org/>) is used to integrate Python's pipeline (Fig. 3) code on Java's platform.



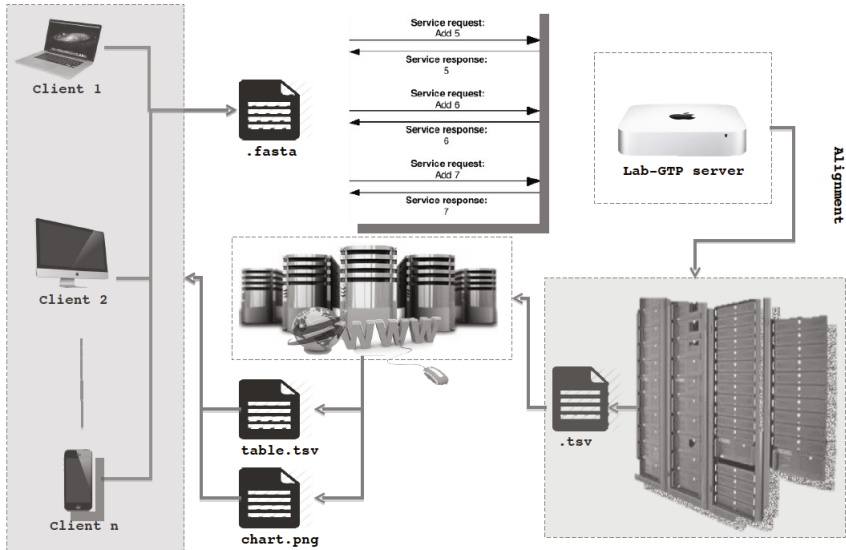
**Fig. 1.** Transcriptator block diagram

## 2.2 System Architecture of Transcriptator Pipeline

Transcriptator pipeline consist of three major components: (i) BLAST analysis, (ii) Gene ontology, (iii) Functional annotation, retrieval and statistical analysis of the data. It requires various levels of computational hardwares (Fig 2). This pipeline is embedded in web application written in Java and Python scripts. The front end user interface of Transcriptator is installed on LAB-GTP server. It helps the user's to submit their queries using our web application interface. The core engine of the pipeline is written in Python, it comprises of the blast analysis as well as different web services for functional annotation analysis from publicly available annotation databases such as DAVID and Quick GO. The core engine is locally installed on a interomics cluster which is connected to LAB-GTP server. For BLAST analysis, ncbi-blast.2.2.23 stand alone package is installed on the cluster. SwissProt and UniProt-trEMBL databases (<http://www.uniprot.org/>) are also installed for BLAST run. DAVID and Quick-GO webservices are installed on the cluster for the faster processing of results. The query of FASTA sequence datasets provided through web application on our web server is directly transferred to our interomics cluster. Local BLAST analysis is carried out on the local cluster implying BLAST X run on locally installed SwissProt and UniProt databases. BLAST results are analysed and top proteins hits id's are used as input for DAVID and QUICK-GO web-services to retrieve functional and gene ontological annotations. The retrieved data is processed and feeded to statistical analysis section of Transcriptator pipeline core engine. The results are provided in the form of graphs and tabular reports, and transferred to the LAB-GTP web server again. From the server, user can access to this information by using the job IDs provided by the server.

## 2.3 Pipeline Implementation

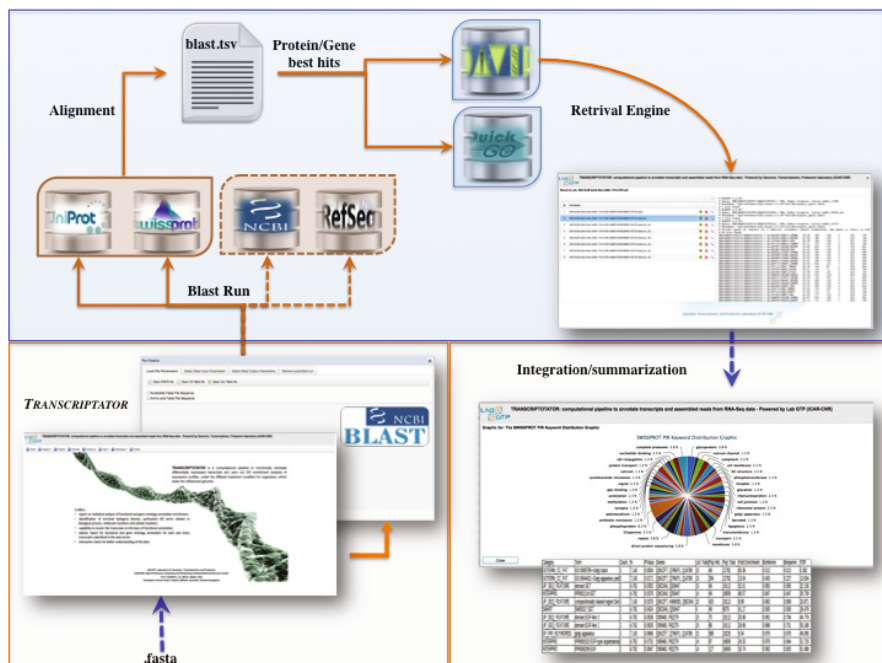
The Transcriptator pipeline is written in Python, bash and R scripts. It implements the web services available for DAVID and QuickGO tools. For DAVID web-services, it utilizes the available Python client source code. The Python client for DAVID web-services, which use light-weight soap client suds-0.4 module [<https://pypi.python.org/pypi/suds>] [6]. For QuickGO web-services, BioServices Python package is implemented in the pipeline [9]. It provides access to QuickGO and a framework to easily implement web service wrappers (based on WSDL/SOAP or REST protocols). In this pipeline, the annotation process comprises of four main parts: (i) finding the best hit in locally installed SwissProt and UniProt-Trembl database; (ii) assignment of functional annotation and gene ontology terms and their enrichment from DAVID; (iii) assignment of GO Slim terms and their analysis from QuickGO; (iv) integration and summarization of retrieved results from DAVID and QuickGO web services. Transcriptotator runs the first step of BLAST search on the local cluster. The second and third steps of pipeline simultaneously run to accelerate the annotation procedure. The last step retrieves the results, processes them and generates the statistical reports in the form of tables and charts.



**Fig. 2.** System architecture of Transcriptator Pipeline

**Identification of Best Hit.** BLASTX program from locally installed ncbi-blast.2.2.23 stand alone package [10] is used (with threshold E-value 0.001) to identify the best hits for query sequences on locally installed SwissProt and UniProt-trEMBL databases (<http://www.uniprot.org/>). The main goal of the first step is to find the similar sequences within SwissProt and UniProt-trEMBL databases for the unannotated query from the user. The output of BLASTX run is an alignment file in a tsv format. The latter, using a bash script, is transformed into the protein list, which is the required input file for DAVID and QuickGO web services.

**Assignment of Functional and Gene Ontology Annotation from DAVID.** Python client source code for DAVID, retrieves the functional and gene ontology annotation for every single transcript in a query data set. These python scripts take the input protein list file from previous step and utilize DAVID database to obtain information in the form of ChartReport, ClusterReport, TableReport and SummaryReport. For a given query data set, Python source code implemented within the Transcriptator pipeline runs with default parameter for DAVID database search to obtain the enrichment statistics for each functional and GO term. ChartReport is an annotation-term-focussed view, which lists annotation terms and their associated genes under study. It also provides the Fischer exact statistics calculated for each annotation term and information about the statistically enriched annotation terms in the query data set. The ClusterReport displays the grouping of similar annotation terms along with their associated genes. The grouping algorithm is based on the hypothesis that similar annotations should have similar gene members.



**Fig. 3.** Transcriptator pipeline: the lower panel boxes respectively show the input/output of the web interface, whereas the upper panel represents the steps of the Transcriptator engine.

**Assignment and Analysis of GO Slim Terms from QuickGO.** Transcriptator employs BioServices module from Python package, which provides access to many bioinformatics web services and a framework to easily implement web service wrappers (based on WSDL/SOAP or REST protocol). BioServices (bioservices.quickgo.QuickGO) is used to investigate the GO slims in the query data set. GO slim terms are the list of GO terms that have been selected from the full set of terms available from the gene ontology projects.

**Processing of Retrieved Annotation.** Both DAVID and QuickGO web services can produce large amount of results for the given query data set. It is not possible for the users to understand and interpret this bulk amount of results in a simple way. For the integration and summarization of retrieved results from web-services, Python and R codes in Transcriptator are implemented to parse the results in simpler format. Transcriptator produces easy to read tables for enrichment analysis of GO and Functional terms, clustering analysis on transcripts and annotation assignment for every single transcript. R scripts are specifically implemented in the pipeline, to generate an interactive chart for the distribution of functional and GO terms such as biological process, molecular function and cellular components associated with the query data set of transcripts.

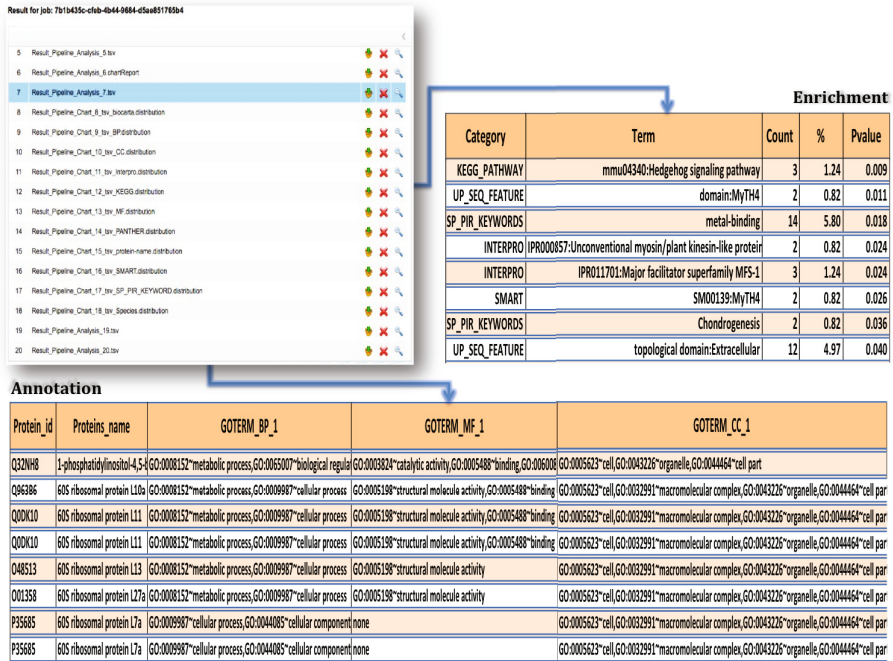
### 3 Results

Transcriptator web application provides a user a friendly interface to input unannotated transcripts or denovo assembled reads from RNA-Seq experiments in multi fasta file format. As DAVID web services limit the analyses to 3000 transcripts at time, our web server also allows a user to input up to 3000 transcripts for annotation. After a successful submission, a unique job ID is generated and provided as an identifier to start the annotation. All annotation results from DAVID and QuickGO are obtained through our server and the user can download them using the associated job ID. The results for single job id comprise of several tables and graphs. The tables are divided into three sections. The first section contains the table with the list of the best hit proteins with E-value from the databases for the corresponding transcript. The second section comprises of the tables generated from the DAVID annotation analysis. It includes four tables: ChartReport (for enrichment analysis); ClusterReport (for clustering analysis); TableReport for functional and GO annotation for every single transcripts in the dataset; SummaryReport with the summary of total annotation for the given query data set. The third section comprises of the table enlisting the assignment of GO slim terms on the transcripts. The pipeline also produces charts related to the distribution of GO terms specifically related to three categories of biological process, molecular function and cellular components, respectively, for the input query data set. For each job concluded, the related annotation results (Fig. 4) will be retained for one week on the Transcriptator server. User can access to this information by using the job IDs provided by the server.

#### 3.1 Case Study

To demonstrate the utility of Transcriptator in biological studies, we have selected a sample dataset (five hundred and forty four unannotated transcripts) of *Hydra vulgaris* transcriptome, downloaded from European Nucleotide Archive (ENA) database (<http://www.ebi.ac.uk/ena/>). These transcripts are specifically differentially expressed in response to cadmium treatment (unpublished data of specific DE transcripts for cadmium treatment). Cadmium is a toxic element. It accumulates in the organisms body and produce pathogenic changes. To study the harmful effects of cadmium accumulation in the body, previously researchers studies the toxicity and chemical stress due to cadmium concentration in non model organism Hydra [11]. They have shown morphological, developmental and physical damage in Hydra due to the presence of high concentration of cadmium in the organism body. To undermine the molecular mechanism of cadmium poisoning in Hydra, we have investigated these cadmium specific differentially expressed transcripts through our pipeline Transcriptator. It annotates these transcripts for all the functional and gene ontology categories and produces results table and graphical charts for functional annotations as well as gene ontology enrichment analysis (Fig 5).

Our results from Transcriptator shows enrichment of Hedgehog signalling pathway and metal binding biological process functional terms with significant



**Fig. 4.** Result files from the Transcriptator. It includes tables for BLAST results, functional and GO enrichment. Graphical chart for GO and functional annotation terms distribution in input data.























































p-value of .009 and .018 respectively (Fig. 4: enrichment table). These results make biological sense as the Hedgehog (Hh) family of secreted signaling proteins plays a crucial role in development and morphogenesis of a variety of tissues and organs in *Hydra vulgaris*. Fig.6 shows biological process terms distribution chart generated by transcriptator for the sample data set. For the given data set. Transcriptator pipeline also provide distribution plots for both molecular functions as well as cellular components.

Molecular functions distribution plots suggests 42.6 percent of sample dataset of Hydra transcripts involved in binding function. It also shows transcription regulator activity (4.6 percent), transporter activity (6.6 percent), catalytic activity (16.7 percent) and molecular transducer activities (11.5 percent) are enriched in these transcripts dataset of Hydra in response to the cadmium toxicity (Fig. 7).

For the given query dataset, Transcriptator pipeline also provide cellular components distribution plot. Distribution of cellular componets associated to differentially expressed transcripts suggests the role of these transcripts in cellular composition. It also provide some information about other cellular components such as synapse, macromolecular complex region, organelles and membrane enclosed regions but unfortunately these cellular compents are not enriched in our query dataset (Fig 8).

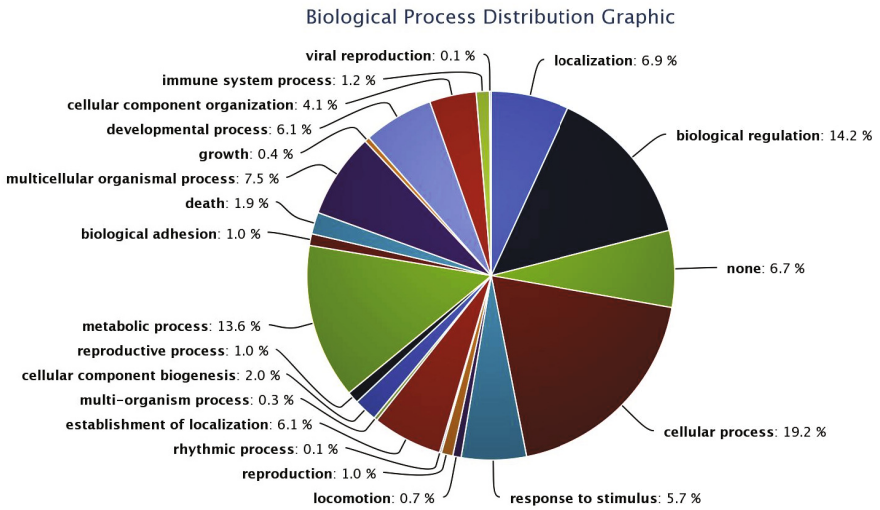


**Result for job: c41a5889-7534-45b0-bf37-ae7a8e39a6ff**

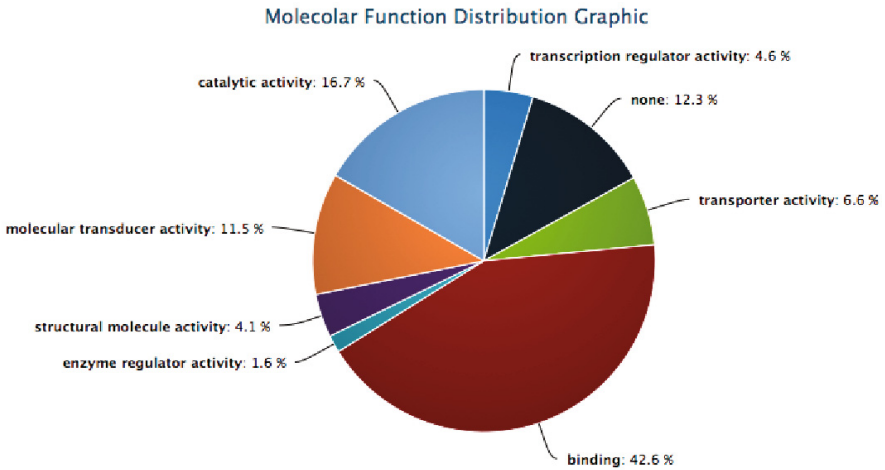
ID	File Name			
1	Processed_fasta_file_1.fasta			
2	Result_BLAST_Analysis_2.tsv			
3	Result_BLAST_Analysis_3.tsv			
4	Result_BLAST_Analysis_4.tsv			
5	Result_BLAST_Analysis_5.tsv			
6	Result_Pipeline_Analysis_Chart_Report_6.chartReport			
7	Result_BLAST_Analysis_7.tsv			
8	Result_Pipeline_Chart_8_tsv_biocarta.distribution			
9	Result_Pipeline_Chart_9_tsv_BP.distribution			
10	Result_Pipeline_Chart_10_tsv_CC.distribution			
11	Result_Pipeline_Chart_11_tsv_Interpro.distribution			
12	Result_Pipeline_Chart_12_tsv_KEGG.distribution			
13	Result_Pipeline_Chart_13_tsv_MF.distribution			
14	Result_Pipeline_Chart_14_tsv_PANTHER.distribution			
15	Result_Pipeline_Chart_15_tsv_protein-name.distribution			
16	Result_Pipeline_Chart_16_tsv_SMART.distribution			
17	Result_Pipeline_Chart_17_tsv_SP_PIR_KEYWORD.distribution			
18	Result_Pipeline_Chart_18_tsv_Species.distribution			

**Fig. 5.** List of results obtained from Transcriptator: tabular results for functional and gene ontology terms enrichments. Graphical distributions plots for GO's terms and functional annotations for a given query data set.

Apart for Gene Ontological terms such as biological process, molecular functions and cellular components, Transcriptator pipeline also provides distribution plots for other different functional terms associated with the query dataset, such as Biocarta pathways, Panther pathways, KEGG pathways, proteins domains such as InterPro, PFAM, SMART. It also provide SP-PIR-keyword distributions. As Transcriptator uses DAVID web services, it can provide each and every relevant functional information related to our query dataset in the the form of distribution plots as well as summary and enrichments table. for example SP-PIR-keyword distribution plot (Fig. 9) shows a large number of keywords are associated to the query dataset, it includes terms like transcription,transducer,alternative splicing, differentiation, dna binding, developmental proteins, g-protein coupled receptor, nucleotide binding, signal and ion transport etc. All these terms associated to the dif-

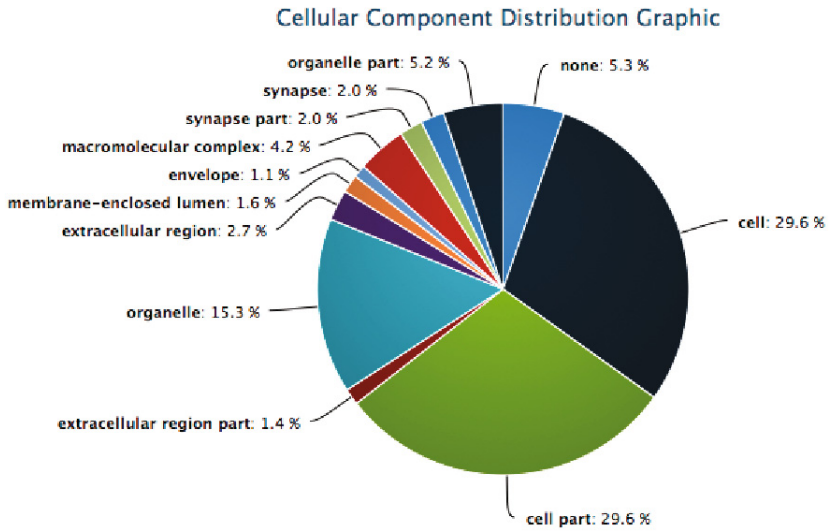


**Fig. 6.** Biological processes distribution in this case study dataset. It shows the significant biological activities, in which these transcripts (case study dataset) are involved. For example biological regulation, cellular process, stimulus response activities and developmental process are enriched within these transcripts.

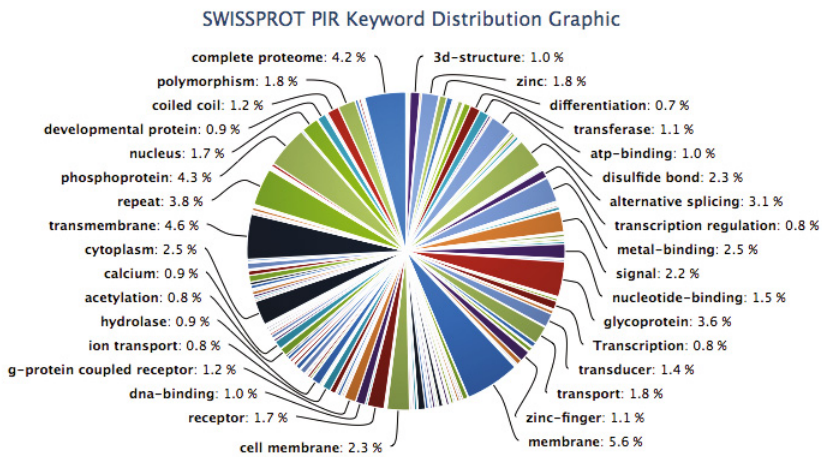


**Fig. 7.** Molecular function distribution in this case study dataset. It shows the significant molecular function activities, in which these transcripts (case study dataset) are involved. For example binding, molecular transducer activity, transcription regulator activity and catalytic activities are enriched within these transcripts.

ferentially expressed transcripts dataset suggest the most possible role of cadmium toxicity on differentiation, reproduction, developmental as well signal transduction processes in *Hydra vulgaris*. Transcriptator pipeline also provides the



**Fig. 8.** Cellular components distribution in this case study data set. It shows the significant cellular components, in which these transcripts (case study dataset) are involved. For example biological regulation, most of the differentially expressed transcripts from the query dataset are associated with cell organization. A small number of transcripts are also involved with structural composition of synapse, macromolecular complex, membrane and cellular organelle but are not statistically significant.



**Fig. 9.** SwissProt-PIR keywords distribution: Graphical representation of all the SP-PIR keywords.

Functional Terms	%	Counts
GENETIC_ASSOCIATION_DB_DISEASE	0,033	8
GENETIC_ASSOCIATION_DB_DISEASE_C	0,033	8
OMIM_DISEASE	0,05	12
COG_ONTOLOGY	0,071	17
PIR_SEQ_FEATURE	0,062	15
SP_COMMENT_TYPE	0,954	230
SP_PIR_KEYWORDS	0,979	236
UP_SEQ_FEATURE	0,975	235
ZFIN_ANATOMY	0,041	10
GOTERM_BP_1	0,809	195
GOTERM_BP_2	0,805	194
GOTERM_BP_3	0,788	190
GOTERM_BP_4	0,768	185
GOTERM_BP_5	0,718	173
GOTERM_BP_ALL	0,809	195
GOTERM_BP_FAT	0,784	189
GOTERM_CC_1	0,859	207
GOTERM_CC_2	0,838	202
GOTERM_CC_3	0,838	202
GOTERM_CC_4	0,822	198
GOTERM_CC_5	0,805	194
GOTERM_CC_ALL	0,859	207
GOTERM_CC_FAT	0,734	177
GOTERM_MF_1	0,809	195
GOTERM_MF_2	0,784	189
GOTERM_MF_3	0,73	176
GOTERM_MF_4	0,672	162
GOTERM_MF_5	0,556	134
GOTERM_MF_ALL	0,809	195
GOTERM_MF_FAT	0,734	177
PANTHER_BP_ALL	0,378	91
PANTHER_MF_ALL	0,378	91
CHROMOSOME	0,834	201
CYTOBAND	0,647	156
ENTREZ_GENE_SUMMARY	0,191	46
HOMOLOGOUS_GENE	0,772	186
OFFICIAL_GENE_SYMBOL	0,934	225
PIR_SUMMARY	0,548	132
SP_COMMENT	0,946	228
GENERIF_SUMMARY	0,402	97
HIV_INTERACTION_PUBMED_ID	0,004	1
PUBMED_ID	0,917	221
ENSEMBL_GENE_ID	0,726	175
ENTREZ_GENE_ID	0,954	230
BBID	0,008	2
BIOCARTA	0,029	7
EC_NUMBER	0,207	50
KEGG_PATHWAY	0,315	76
PANTHER_PATHWAY	0,237	57
REACTOME_PATHWAY	0,025	6
BLOCKS	0,602	145
COG_NAME	0,071	17
INTERPRO	0,971	234
PANTHER_FAMILY	0,705	170
PANTHER_SUBFAMILY	0,593	143
PFAM	0,963	232
PIR_SUPERFAMILY	0,51	123
PRINTS	0,423	102
PRODOM	0,232	56
PROFILE	0,589	142
PROSITE	0,78	188
SCOP_CLASS	0,033	8
SCOP_FAMILY	0,033	8
SCOP_FOLD	0,033	8
SCOP_SUPERFAMILY	0,033	8
SMART	0,515	124
SSF	0,315	76
TIGRFAMS	0,116	28
BIND	0,162	39
DIP	0,071	17
HIV_INTERACTION	0,004	1
HIV_INTERACTION_CATEGORY	0,004	1
MINI	0,158	38
NCICB_CAPATHWAY_INTERACTION	0,021	5
REACTOME_INTERACTION	0,017	4
UCSC_FBBS	0,191	46
CGAP_EST_QUARTILE	0,133	32
CGAP_SAGE_QUARTILE	0,129	31
GNF_U133A_QUARTILE	0,12	29
PIR_TISSUE_SPECIFICITY	0,241	58
UNIGENE_EST_QUARTILE	0,145	35
UP_TISSUE	0,722	174

**Fig. 10.** Summary table: it provides all the functional and GO terms associated with the given query dataset.

Category	Term	Count	%	Pvalue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	mmu04340:Hedgehog sig	3	1.244	0.0095	17	54	5738	18.75	0.326	0.326	8.5
UP_SEQ_FEATURE	domain:MyTH4	2	0.829	0.0114	47	4	16021	170.4	0.872	0.872	13.19
SP_PIR_KEYWORDS	metal-binding	14	5.809	0.0183	48	2682	17854	1.941	0.811	0.811	18.28
INTERPRO	IPR000857:Unconvention	2	0.829	0.0240	49	9	17763	80.55	0.961	0.961	24.81
INTERPRO	IPR011701:Major facilitat	3	1.244	0.0241	49	89	17763	12.21	0.962	0.805	24.86
SMART	SM00139:MyTH4	2	0.829	0.0263	28	9	9131	72.46	0.617	0.617	21.16
SP_PIR_KEYWORDS	Chondrogenesis	2	0.829	0.0362	48	14	17854	53.13	0.963	0.810	33.0
UP_SEQ_FEATURE	topological domain:Extrac	12	4.979	0.0404	47	2174	16021	1.881	0.999	0.975	39.84
INTERPRO	IPR008085:Thrombospon	2	0.829	0.0423	49	16	17763	45.31	0.996	0.855	39.77
UP_SEQ_FEATURE	topological domain:Cytop	14	5.809	0.0451	47	2780	16021	1.716	0.999	0.936	43.32
SP_PIR_KEYWORDS	zinc	10	4.149	0.0548	48	1886	17854	1.972	0.993	0.816	45.87
UP_SEQ_FEATURE	domain:TSP type-1 5	2	0.829	0.0567	47	20	16021	34.08	0.999	0.923	50.74
SP_PIR_KEYWORDS	osteoogenesis	2	0.829	0.0563	48	22	17854	33.81	0.994	0.728	46.80
INTERPRO	IPR001111:Transforming	2	0.829	0.0578	49	22	17763	32.95	0.999	0.864	50.21
UP_SEQ_FEATURE	transmembrane region	18	7.468	0.0602	47	4113	16021	1.491	0.999	0.891	53.42
SP_PIR_KEYWORDS	alternative splicing	18	7.468	0.0603	48	4481	17854	1.494	0.996	0.674	49.22
UP_SEQ_FEATURE	domain:TSP type-1 4	2	0.829	0.0640	47	23	16021	29.64	0.999	0.861	55.70
INTERPRO	IPR001879:GPCR_family	2	0.829	0.0705	49	27	17763	26.85	0.999	0.859	57.5
UP_SEQ_FEATURE	domain:TSP type-1 3	2	0.829	0.0747	47	27	16021	25.24	0.999	0.862	61.56
SP_PIR_KEYWORDS	glycoprotein	15	6.224	0.0759	48	3600	17854	1.549	0.999	0.694	57.67
SP_PIR_KEYWORDS	cell membrane	9	3.734	0.0766	48	1713	17854	1.954	0.999	0.641	58.00
SMART	SM00008:FormR	2	0.829	0.0769	28	27	9131	24.15	0.944	0.763	51.04
INTERPRO	IPR000203:GPS	2	0.829	0.0830	49	32	17763	22.65	0.999	0.855	63.74
INTERPRO	IPR015615:Transforming	2	0.829	0.0830	49	32	17763	22.65	0.999	0.855	63.74
UP_SEQ_FEATURE	domain:GPS	2	0.829	0.0853	47	31	16021	21.99	0.999	0.864	66.64
UP_SEQ_FEATURE	domain:PH 2	2	0.829	0.0879	47	32	16021	21.30	0.999	0.839	67.80
SMART	SM00303:GPS	2	0.829	0.0905	28	32	9131	20.38	0.967	0.679	57.11
UP_SEQ_FEATURE	domain:PH 1	2	0.829	0.0906	47	33	16021	20.65	0.999	0.817	68.92
UP_SEQ_FEATURE	zinc finger region:RING-t	3	1.244	0.0907	47	176	16021	5.810	0.999	0.787	68.99
INTERPRO	IPR017948:Transforming	2	0.829	0.0953	49	37	17763	19.59	0.999	0.853	69.06
INTERPRO	IPR001839:transforming	2	0.829	0.0953	49	37	17763	19.59	0.999	0.853	69.06
UP_SEQ_FEATURE	domain:TSP type-1 2	2	0.829	0.0984	47	36	16021	18.93	0.999	0.786	72.1
UP_SEQ_FEATURE	domain:TSP type-1 1	2	0.829	0.0984	47	36	16021	18.93	0.999	0.786	72.1

**Fig. 11.** Enrichment Chart: tabular results for functional and gene ontology terms enrichments. It provides the statistical P-values for the better interpretation of the results

summary information of all the functional and GO terms associated with the queried dataset along with the percentage and counts of transcripts. this information helps the user to undermine the associated functionalities for the given transcripts.

Through Transcriptator pipeline, for the given query dataset of Hydra transcripts, the complete tabular results for enriched functional and GO terms are obtained (Fig 11). It shows the different functional terms, such as pathways, protein domains, SP-PIR keywords which are enriched in the given query data set. It also provide the counts of transcripts and enriched P-values. Transcriptator pipeline also provide corrected P-values after Bonferroni and Benjamini correction. False discovery rate and fold enrichments information is also provided for the given functional terms associated with the transcripts dataset. It enables user to determine the statistical significance for the functionalities associated to their input transcripts dataset for the better biological interpretation of their transcriptomic data.

## 4 Conclusion

Transcriptator is a modular pipeline, which provides flexibility to user to carry out functional annotation of transcript's data. It allows users to choose two distinct types of web services for annotation purposes, as well as different BLAST databases for BLAST run. All these options help users to optimize their results, according to their needs. It provides the enrichment score for the functional

terms and reports each and every annotation present in the given data set in the form of tables and interactive charts. In future, we will work on the addition of more modular functionalities and options in the pipeline, for both BLAST searches and annotation analysis.

**Acknowledgements.** We would like to thank the INTEROMICS flagship project, PON02-00612-3461281 and PON02-00619-3470457 for the funding support. Mario Guarracino work is conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

## References

1. Steijger, T., et al.: Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184 (2013)
2. Huang, D.W., et al.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4, 44–57 (2009)
3. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37(1), 1–13 (2009)
4. Binns, D., et al.: QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 15, 25(22), 3045–3046 (2009)
5. Chen, T.W., et al.: FastAnnotator- an efficient transcript annotation web tool. *BMC Genomics* 13(Suppl. 7), S9 (2012)
6. Jiao, X., et al.: DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28(13), 1805–1806 (2012)
7. Wang, X., Cairns, M.J.: Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics* 14(Suppl. 5), S16 (2013)
8. Nagaraj, S.H., et al.: ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. *Nucleic Acids Res.* 35(Web Server issue), 143–147 (2007)
9. Cokelaer, T., et al.: BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* 29, 3241–3242 (2013)
10. Altschul, S.F., et al.: Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990)
11. Karntanut, W., Pascoe, D.: The toxicity of copper, cadmium and zinc to four different Hydra (Cnidaria: Hydrozoa). *Chemosphere* 47, 1059–1064 (2002)