GOmir: a stand-alone application for human microRNA target analysis and gene ontology clustering

Pantelis Zotos*, Georgios Papachristoudis*, Maria G. Roubelakis, Ioannis Michalopoulos, Kalliopi I. Pappa,

Nikolaos P. Anagnou and Sophia Kossida.

*The authors contributed equally to the work.

Abstract— **MicroRNAs (miRNAs) are single-stranded RNA molecules of about 20–23 nucleotides length found in a wide variety of organisms. MiRNAs regulate gene expression, by interacting with target mRNAs at specific sites in order to induce cleavage of the message or inhibit translation. Predicting or verifying mRNA targets of specific miRNAs is a difficult process of great importance. GOmir is a novel stand-alone application consisting of two separate tools: JTarget and TAGGO. JTarget integrates miRNA target prediction and functional analysis by combining the predicted target genes from TargetScan, miRanda, RNAhybrid and PicTar computational tools and also providing a full gene description and functional analysis for each target gene. On the other hand, TAGGO application is designed to automatically group gene ontology annotations, taking advantage of the Gene Ontology (GO), in order to extract the main attributes of sets of proteins. GOmir represents a new tool incorporating two separate Java applications integrated into one stand-alone Java application. GOmir (by using up to four different databases) introduces, for the first time, miRNA predicted targets accompanied by (a) full gene description, (b) functional analysis and (c) detailed gene ontology clustering. Additionally a reverse search initiated by a potential target can also be conducted. GOmir can freely be downloaded from http://bioacademy.gr/bioinformatics/projects/GOmir.**

I. INTRODUCTION

MicroRNAS (miRNAs) are 20- to 23- nucleotide long single stranded RNAs that post-transcriptionally regulate gene expression [1, 2]. MiRNAs act as translation inhibitors of mRNA into protein and promote mRNA degradation. In this way, miRNAs play important role in various cell processes such as proliferation, differentiation, apoptosis, development, cancer and various other diseases [1, 2] and thus represent potential targets for therapeutic applications.

The biogenesis of miRNAs is a complicated process involving two different cellular compartments [3]. First, in the nucleus, a primary miRNA (pri-miRNA) is transcribed from the genomic DNA by RNA polymerase II. The size of this primary product varies from 100- to 1000- nucleotides in length. Then, the pri-miRNA is truncated by Drosha and DGCR8 to form a hairpin loop precursor called pre-miRNA [3]. The 60-70 nucleotide long pre-miRNA is loaded to Exportin 8 and Ran-GTP in order to be exported into the cytoplasm. A mature miRNA (20-23 nucleotides) is then released by the RNAse III endonuclease complex including Dicer and trans-activator RNA (tar)-binding protein TRBP [3]. The mature miRNA then inhibits translation of a miRNA into a protein by imperfect base pairing to one or more mRNA sequences [1, 4]. The identification of human miRNAs and their respective targets is of great importance and involves both computational and experimental approaches. Prediction servers such as TargetScan [5], miRanda [6], RNAhybrid [7] and PicTar [8] give information for the miRNA-target interactions. Recent reports have described correlated computational expression of miRNA and their target mRNAs and proteins giving a detailed functional description of the latest [4, 9]. Herein, we describe GOmir, a new stand-alone application for human miRNAs target prediction and ontology clustering, consisting of two different components, JTarget and TAGGO. JTarget combines the data from four different databases (TargetScan, miRanda, RNAhybrid and PicTar), whereas TAGGO gives detailed assignments from Gene Ontology (GO) resources to gene products. TAGGO uses one of the most reliable biological ontologies, the Gene Ontology, the main goal of which is to provide a well structured, precisely defined and controlled vocabulary for describing the roles of genes and gene products in any organism. Thus, GOmir serves as a reliable tool for miRNA target prediction and more interestingly provides assignments from GO resources for these gene products, exploring in this way the functional aspects of miRNAs in more detail.

II. METHODS

We will describe the design and implementation of both tools JTarget and TAGGO.

A. JTarget

1) Data Acquisition

Data derived from human miRNA target predicting tools, such as TargetScan, miRanda, RNAhybrid and PicTar were used. The TargetScan database was obtained from the TargetScan website (the 4.2 release of April 2008, http://www.targetscan.org/). Concerning the miRanda tool, the latest up-to-date data were downloaded from miRBase web site (http://microrna.sanger.ac.uk/index.shtml). Similarly, the data from RNAhybrid database were retrieved from the mirnamap website (http://mirnamap.mbc.nctu.edu.tw/). Finally, the PicTar data were obtained from the UCSC genome browser database (http://genome.ucsc.edu/). The database files were treated, in order to obtain only the human target genes. For gene description and functional analyses, three database files were downloaded from the DAVID Bioinformatics database (http://david.abcc.ncifcrf.gov/home.jsp), in order to implement the "Find gene description" and "Find gene function" applications and correlate in this way each gene product with a description and a function analysis, respectively.

2) Data Integration

The database files from the four miRNA target prediction tools were truncated to the human related information, in order to have the minimum size and all the human miRNAs were paired with the respective targets. The TargetScan database file contains miRNA families, instead of individual miRNAs. Therefore, the miRNA families file corresponding to the respective miRNAs was downloaded as well. Different gene ID systems (Refseq ID, Gene symbol, Ensembl ID) are used among different databases. In order to correlate the data among different data sets the NCBI website (http://www.ncbi.nlm.nih.gov/) and the DAVID Bioinformatics Database (http://david.abcc.ncifcrf.gov/home.jsp) were used. The downloaded files, from the DAVID database for the "Find genes description" and "Find genes functions" procedures, contained pairs of DAVID ID number/ genes symbol, DAVID ID number/ genes description and DAVID ID number/ genes functions and were minimized to the human related information.

For the functionality and performance of the application we decided to create a database with all the necessary files, which were described above. We used SQLite, a software library that implemented a self-contained, serverless, zero-configuration, transactional SQL database engine and is ideal for internal databases used for distributable, stand-alone application (http://www.sqlite.org/). We imported the information from our data files in a SQLite database file, necessary for JTarget

functionality and is downloaded along with the entire application installation package.

3) Functionality and implementation

JTarget main goal is to find the target genes of a given miRNA derived from 1 to 5 databases and compare any possible combination of the results from these databases. The available databases are TargetScan, miRanda, RNAhybrid, PicTar-4way and PicTar-5way. Following the insertion of a specific miRNA name, 3 possible scenarios are given. The user can obtain the resulted target genes from: (a) only one database, (b) 2 or more databases or (c) all combined databases. The miRNA target genes search within a single database is implemented by executing a "SELECT microRNA, target FROM database_name WHERE microRNA=microRNA_name" sql query into the database. The common targets from several database tools are found by performing inner joins among the results from the respective "SELECT" statements.

JTarget comprises some more functionalities besides the miRNA common target genes prediction. After target gene selection for a given miRNA, the user can search for a description of these target genes or for their functions. These two options are implemented by executing "SELECT" queries into the entire database in order to correlate each gene with a description and/or functions, respectively. In addition, it is possible to perform a reverse search. The user may enter a gene symbol (e.g. STAG2 for STROMAL ANTIGEN 2) and obtain the common miRNAs of the selected databases for the given gene.

Finally, the JTarget tool is connected to the TAGGO through a button named "TAGGO", which enables the clustering of the genes. A temporary file is constructed from the output file from a miRNA target search, and then used to the TAGGO tool.

B. TAGGO

Gene Ontology (GO) is divided into three ontology aspects which yield information common to all living organisms. Molecular Function (MF) and Cellular Component (CC) aspects answer the questions of what a gene product does and where its active form can be found, whereas the Biological Process (BP) aspect clarifies the biological objective of a gene product. Each ontology aspect is structured as a Direct Acyclic Graph (DAG), a graph with no cyclic paths (no loops) with its nodes representing the ontology terms (and their intrinsic properties) and its edges the relations between nodes. Each GO term has a unique identifier which is used as a database cross-reference in the collaborating databases [10]. Each gene product-GO term pair is followed by an Evidence Code (EC) which indicates how an annotation to a particular term is supported. There are fourteen different ECs. The higher the specificity of a term, the lower its level inside the ontology hierarchy is, and *vice versa*. Proteins are often annotated with terms of medium or low level in ontology. This provides a huge amount of information that is misleading when the aim is to pin-point the main characteristics and functions of a protein or a protein set. To obtain a more global

view of the attributes of a protein or a protein set, a way to assign more general terms to proteins, is needed. Finding more general categories for the function and localization of a protein is equivalent of tracking the most general terms of GO which are relevant to its annotated GO terms, as more generic terms (those high in the ontology) mainly serve as abstractions which demonstrate the broader role of their children.

1) Algorithm

TAGGO implements general terms in the GO DAG structure and automatically produces biologically meaningful results. A method to estimate the specificity of a term is the evaluation of its Information Content (IC). In Algorithmic Information Theory, the information content of an individual object is a measure of the degree of difficulty to define or describe that object [11]. In other words, high information content implies more intense effort to process an object. In biological terms, the higher the information content of a GO term, the more specific this term is and vice versa.

To confine this theoretical definition into a mathematical formula, it is necessary to consider that the times a term occurs denote how general this term is. It is not even necessary to encounter the term itself but any of its children: According to the *True Path Rule*, a rule imposed in order to ensure the validity of GO entries, the pathway from a child term to its top-level parent(s) must always be true [10]. In other words, a term holds all the attributes of its ancestors and can be considered one of them.

Measurement of the degree of specificity of a term is complicated by the fact that the local density of GO terms and the length of branches vary. Furthermore, "leaves" should contain the same IC, as they provide the most detailed descriptions [12].

Formalistically:

$p(c)$, the probability of a term, is defined as the frequency of encountering this term:

$$p(c) = \frac{n_c}{n_r}$$

where n_c is the number of times this term (or any of its children) occur and n_r is the number of times any term (or the root term) occurs [13]. The information content $IC(c)$ of this term is defined as the negative logarithm of its probability:

$$IC(c) = -\log(p(c)) = -\log(\frac{n_c}{n_r})$$

As any leaf can occur only once, its probability is:

$$p(leaf) = \frac{1}{n_r}$$

and its Information Content:

$$IC(leaf) = -\log(p(leaf)) = -\log(\frac{1}{n_r})$$

To normalise the Information Content of a term, its IC is divided by that of a leaf:

$$IC_{normal}(c) = \frac{IC(c)}{IC(leaf)} = \frac{\log(p(c))}{\log(p(leaf))} =$$

$$= \frac{\log(\frac{n_c}{n_r})}{\log(\frac{1}{n_r})} = 1 - \frac{\log(n_c)}{\log(n_r)}$$

$IC_{normal}(c)$, the Normalised Information Content of a term, ranges from 0 (root) to 1 (leaf).

2) Functionality

The main input file of TAGGO is a list of proteins that is experimentally produced by e.g. a large scale analysis. SwissProt accession number, gene symbol or International Protein Index (IPI) can be used to identify each protein. To load the GO structure, a GO file is used as input. OBO v.1.0, OBO v.1.2, GO (which is deprecated), OBO-XML or RDF formats are supported. To map gene products to GO terms, the organism of origin must be selected. That triggers the program to load the corresponding to the selected species GO annotation (GOA) file. Each GOA entry provides information about the database which contributes to this annotation, the date of the annotation, the object which is annotated, its synonym, its type (e.g. gene, transcript, protein), its assigned GO term, the ontology, where this term belongs to and evidence about the credibility of this annotation. Users are strongly advised to use the latest GO and GOA files which can be downloaded from the GO FTP site (ftp://ftp.geneontology.org/pub/go/). To increase versatility and robustness, the user has the opportunity to exclude GOA entries supported by less reliable Evidence Codes (ECs). Thus, the output file may only hold the GO annotations of the input proteins supported by well established methods. To exclude very generic terms from the classification, non-desired terms can be specified and normalised information content threshold for the three aspects can be set (default values are 4% or 0.04). Finally, the directory, where the results will be stored is chosen and all data are submitted.

When the program starts running, a file which contains the GO terms of each protein for all three aspects, is created. Then, the protein dataset is categorised into general GO terms, as follows: all parents of each term are found (considering all possible pathways to the root) and sorted according to ascending information content. The most general term which does not belong to the non-desired terms of the corresponding aspect is considered a category, unless all of the ten most general parents belong to the user-specified non-desired terms; in that case, the term is classified as "NO ENTRY" category. The proteins with their assigned GO categories are gathered and duplicated categories for a given protein are removed. The output is visualised in pie and bar charts which show the

percentage of each GO category on the given protein dataset, in all GO aspects. Moreover, Venn lists for all aspects are created to indicate the overlaps of GO categories for the given protein dataset. These lists can be imported in VennMaster [14] to create Venn diagrams. These three types of output indicate how many proteins share a common GO category.

C. GOmir GUI implementation and prerequisites

Both tools were developed in Java Programming Language. We used the widget toolkit for Java, Swing, in order to develop the graphical user interface. As far as the JTarget database implementation is concerned, we selected the SQLite SQL database engine, which does not need any server to be installed and is very compact. In addition, Spring Framework and JFreeChart library were used for the implementation of TAGGO chart functionality. GOmir can be installed in any Microsoft Windows operation system with Java Runtime Engine 1.5.0 (JRE 5.0) (http://www.java.com/) pre installed.

III. USER INTERFACE

JTarget contains text fields, buttons and one central text area for the interaction with the user (Figure I). To find the target genes of a miRNA from a single database, the "Search by miRNA" option has to be selected and a miRNA name (e.g. miR-21) should be given. Then, a database (e.g. TargetScan) should be marked and the results are presented in the central text area by clicking the "Find Targets" button. The common results from 2 or more databases can be obtained by marking the desirable databases and clicking the "Compare database results" button. The available databases are: TargetScan, miRanda, RNAhybrid, PicTar-4way and PicTar-5way. In this study, for the first time, the RNAhybrid database sets were introduced in a computational program.

To obtain the common target genes results for a given miRNA from all the databases (TargetScan, miRanda, RNAhybrid and PicTar-4way), only the "Compare All" button should be clicked without the necessity to select any databases (Figure II). The databases, that give no results for a given miRNA, are excluded from the procedure and the common target genes from the remaining databases are provided.

The JTarget includes datasets from PicTar 4-way and the 5-way databases. Both databases contain target prediction for all human miRNAs based on conservation in mammals. The PicTar 4-way includes human, mouse, rat, dog species datasets, whereas the 5-way includes also and chicken species datasets. However, the insertion of PicTar 5-way reduces the common target genes results to a significant degree. Therefore, we decided to exclude this database from the calculation of the common target genes at the "Compare All" procedure. The number of the predicted targets by searching one or more databases for miR-21, miR-31, miR-221 and miR-222 is shown in TABLE I.
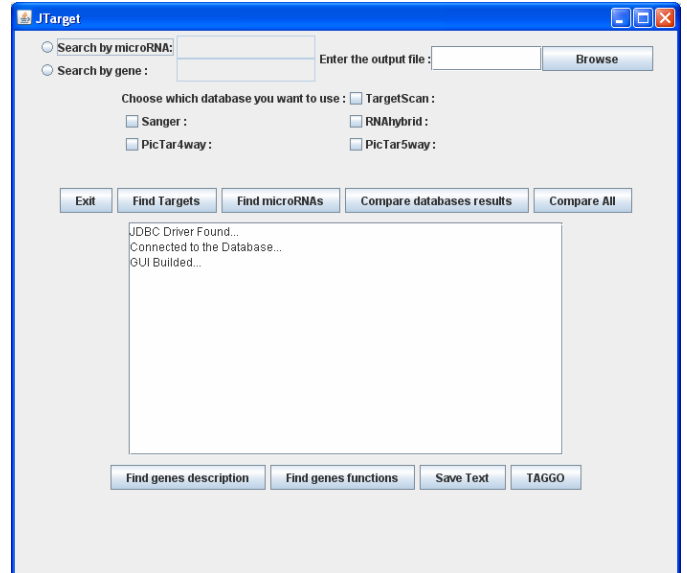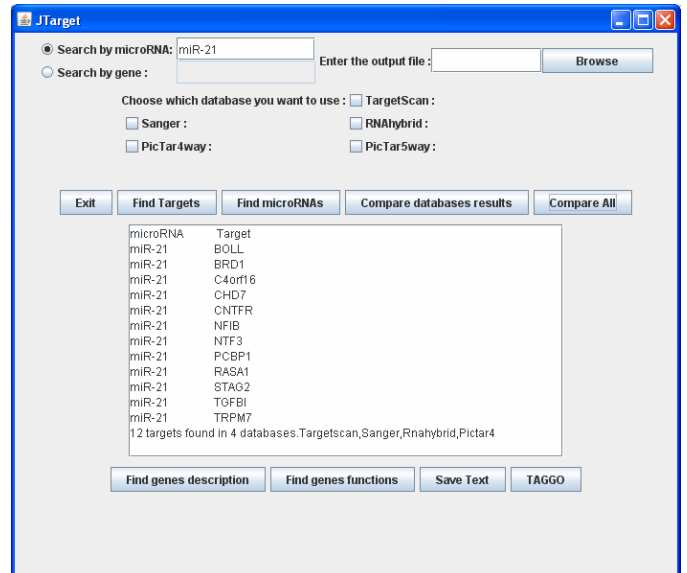

Figure I: JTarget main menu.

Figure II: The target genes of miR-21 as given by the procedure "Compare All" (4 databases).

TABLE I
NUMBER OF COMMON TARGETS FOUND FOR SEVERAL miRNAs FROM 1, 2, 4 OR 5 DATABASES

miRNAs	TargetScan	TargetScan and miRanda	Compare All (4)	Five databases
miR-21	186	44	12	6
miR-31	192	20	7	0
miR-221	251	49	11	5
miR-222	251	40	10	7

The results can be saved either before the search is accomplished, by filling the "Enter the output file" text field, or by choosing "Save Text" after the completion of the search.

Afterwards, the user may find the descriptions of the target genes results just by clicking the "Find genes description" button. Each gene followed by its description is presented in the text area. In addition, the functions of each target gene can be found when the "Find genes functions" is clicked. The information about the genes description and functions is obtained by the DAVID Bionformatics database. In Table II the descriptions of the common genes from 4 databases predicted for the miR-21 are described.

TABLE II

DESCRIPTION OF THE COMMON TARGET GENES OF mir-21 FROM 4 DATABASES

Gene	Description
BOLL	BOL, BOULE-LIKE (DROSOPHILA)
BRD1	BROMODOMAIN CONTAINING 1
C4orf16	CHROMOSOME 4 OPEN READING FRAME 16
CHD7	CHROMODOMAIN HELICASE DNA BINDING PROTEIN 7
CNTFR	CILIARY NEUROTROPHIC FACTOR RECEPTOR
NFIB	NUCLEAR FACTOR I/B
NTF3	NEUROTROPHIN 3
PCBP1	POLY(RC) BINDING PROTEIN 1
RASA1	RAS P21 PROTEIN ACTIVATOR (GTPASE ACTIVATING PROTEIN RAS P21)
STAG2	STROMAL ANTIGEN 2
TGFBI	TRANSFORMING GROWTH FACTOR, BETA-INDUCED, 68KDA
TRPM7	TRANSIENT RECEPTOR POTENTIAL CATION CHANNEL, SUBFAMILY M, MEMBER 7

In TABLE III the functions of the common target genes predicted by the JTarget "Find genes functions" procedure for miR-21 from all 5 databases are shown. The blank cells next to certain genes mean that these genes are not contained in DAVID's gene function database.

TABLE III

FUNCTIONS OF THE COMMON TARGET GENES OF MIR-21 FROM 5 DATABASES

Gene	Function
NFIB	NFIB is capable of activating transcription and replication.
C4orf16	
RASA1	RASA1 is an inhibitory regulator of the Ras-cyclic AMP pathway.
STAG2	STAG2 is a component of cohesin complex, required for the cohesion of sister chromatids after DNA replication.
NTF3	NTF3 promotes the survival of visceral and proprioceptive sensory neurons.
CHD7	

Once a search is completed, a gene ontology clustering may be performed by clicking the "TAGGO" button. A temporary file is created from the output file of a JTarget search and is then used as an input file for TAGGO search. TAGGO interface consists of 5 steps (Figure III). First, the user has to provide the path of the input file containing the proteins. When TAGGO is initiated through JTarget, the first step of TAGGO is done automatically, as the input file is provided directly by JTarget. The second step consists of the selection of the gene ontology file and its format. At the third step, the user is asked to choose the organism for which the clustering is going to be performed. Afterwards, the user may select the Evidence Codes to be included during the annotation process (all of them are included by default), may exclude some non-desired terms or may set a normalised information content threshold for the three aspects. The fifth step is to provide the program with an output directory. The output of the whole process is a directory called "Results", which includes the charts for the visualization of the output, three Venn lists for each GO aspect and some text files with information about the process and its results. For example, according to TAGGO results for the common target genes for miR-21 from TargetScan and miRanda, the most dense GO term for Cellular Component GO aspect is GO:0005562 which corresponds to intracellular (Figure IV) and for Molecular Function GO aspect is GO:0005488 which corresponds binding functions, respectively (Figure V).
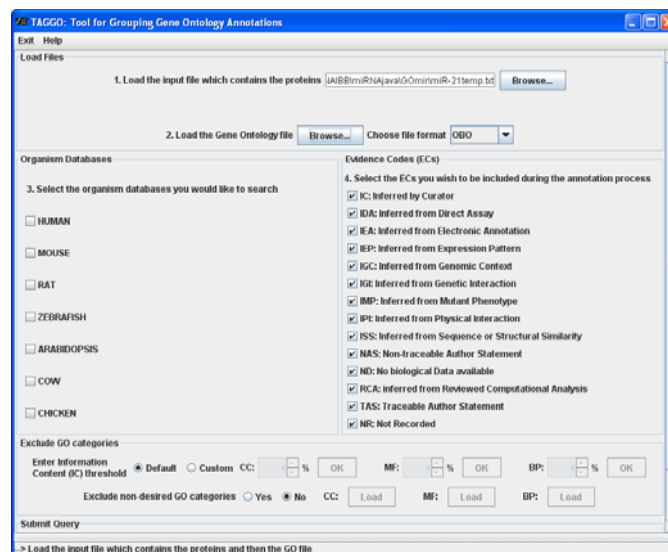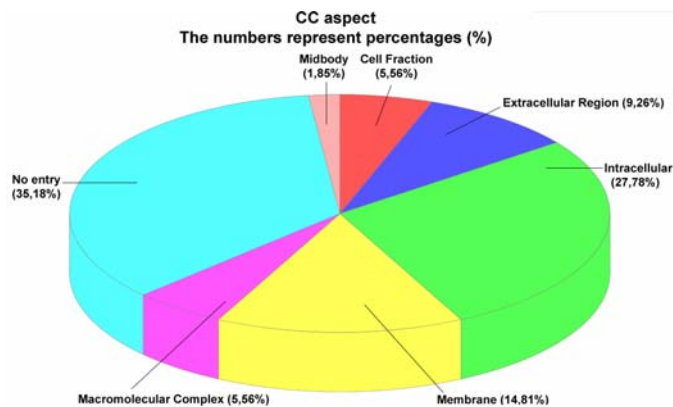

Figure III: TAGGO main menu

CC aspect
The numbers represent percentages (%)

Figure IV: Subcellular distribution of the common target genes for miR-21 as derived from TargetScan and miRanda datasets.

MF aspect
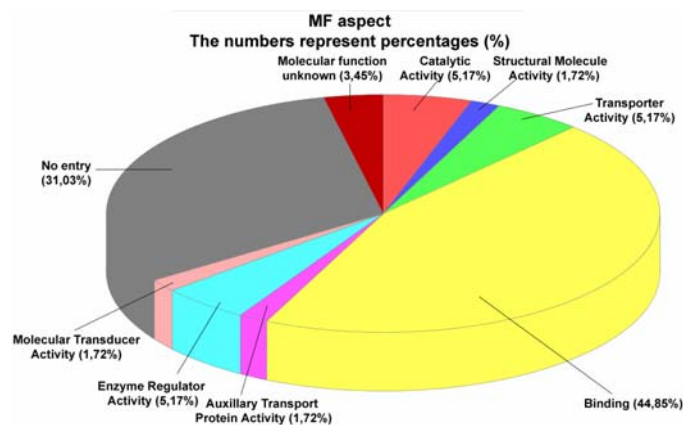**The numbers represent percentages (%)**

Figure V: Functional classification of the common target genes for miR-21 as derived from TargetScan and miRanda datasets.

A reverse search is possible to be performed in JTarget, using the "Search by gene" option, in order to find the complementary miRNAs for the given gene. Searches for common miRNAs from 2 to 5 databases are available as well.

IV. CONCLUSIONS

GOmir, is a novel stand-alone application designed to elucidate the human miRNA interactions with the respective targets by using the data sets retrieved by four different computationally miRNA prediction databases, increasing in this way the validity of the results. In this study, RNAhybrid database was inserted for the first time in a computational tool that combines the results from different miRNA databases. The validity of the computational predicted targets is confirmed by recent experimental studies for certain miRNAs. For example, GOmir indicated NFI-A as possible target for miR-223 comparing 4 out of 4 databases (Sanger, TargetScan, RNAhybrid and PicTar-4way). Further experimental studies by Fazi et al. confirmed this prediction and showed that miR-223 plays a crucial role during granulopoiesis by downregulating NFI-A [15]. GOmir provides a detailed gene

description of the predicted targets accompanied by a function analysis. A reverse search initiated by a potential target can also be performed to find the predicted interacting miRNAs. In a second next step, a detailed gene ontology clustering, including all the respective graphic charts and diagrams for the predicted targets are provided by the TAGGO module of GOmir. In this way, any group of human miRNAs and respective targets can be analysed with functional and ontology information provided, easily, in a short period of time and without using a web-based interface.

V. ACKNOWLEDGMENTS

VI. REFERENCES

[1]   Buchan, J.R. and R. Parker, *Molecular biology. The two faces of miRNA.* Science, 2007. **318**(5858): p. 1877-8.

[2]   Chen, J.F., et al., *The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation.* Nat Genet, 2006. **38**(2): p. 228-33.

[3]   Song, L. and R.S. Tuan, *MicroRNAs and cell differentiation in mammalian development.* Birth Defects Res C Embryo Today, 2006. **78**(2): p. 140-9.

[4]   Megraw, M., et al., *miRGen: a database for the study of animal microRNA genomic organization and function.* Nucleic Acids Res, 2007. **35**(Database issue): p. D149-55.

[5]   Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

[6]   John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.

[7]   Rehmsmeier, M., et al., *Fast and effective prediction of microRNA/target duplexes.* RNA, 2004. **10**(10): p. 1507-17.

[8]   Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.

[9]   Nam, S., et al., *miRGator: an integrated system for functional annotation of microRNAs.* Nucleic Acids Res, 2008. **36**(Database issue): p. D159-64.

[10]   The Gene Ontology Consortium, *Creating the gene ontology resource: design and implementation.* Genome Res, 2001. **11**(8): p. 1425-33.

[11]   Chaitin, G., *Algorithmic Information Theory*, in *Encyclopedia of Statistical Science*. 1982, Wiley: New York. p. 38-41.

[12]   Bérard, S., L. Tichit, and C. Herrmann, *ClusterInspector: a tool to visualize ontology-based relationships between biological entities*, in *Actes des Journées Ouvertes Biologie Informatique Mathématiques*. 2005: Lyon. p. 447-457.

[13]   Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.

[14]   Kestler, H.A., et al., *Generalized Venn diagrams: a new method of visualizing complex genetic set relations.* Bioinformatics, 2005. **21**(8): p. 1592-5.

[15]   Fazi, F., et al., *A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis.* Cell, 2005. **123**(5): p. 819-31.