Frontiers in **Bioengineering**

Methods Article
14 December 2020

Comprehensive reconstruction and visualization of non-coding regulatory networks

Vincenzo Bonnici [1], **Francesco Russo** [2,3], **Nicola Bombieri** [1], **Alfredo Pulvirenti** [4,*,τ] **and Rosalba Giugno** [4,*,τ]

[1] *Department of Computer Science, University of Verona, Verona, Italy*
[2] *Laboratory of Integrative Systems Medicine (LISM), Institute of Informatics and Telematics (IIT) and Institute of Clinical Physiology (IFC), National Research Council (CNR), Pisa, Italy*
[3] *Department of Computer Science, University of Pisa, Pisa, Italy*
[4] *Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy*

Correspondence*:
Rosalba Giugno
Department of Clinical and Molecular Biomedicine, University of Catania, Viale Andrea Doria 6, Catania, 95125, Italy, giugno@dmi.unict.it

Alfredo Pulvirenti
Department of Clinical and Molecular Biomedicine, University of Catania, Viale Andrea Doria 6, Catania, 95125, Italy, apulvirenti@dmi.unict.it

Bioinformatics of Non-Coding RNAs with Applications to Biomedicine: Recent Advances and Open Challenges

ABSTRACT

Keywords: microRNAs, lncRNAs, non-coding RNAs, networks, cytoscape, gene expression

1 INTRODUCTION

After the sequencing of the human genome it became evident that only 20,000 genes are protein-coding, while over 98% of all genes are untranslated non-protein-coding RNAs (ncRNAs) (**Consortium et al.** (2012)). During the last years thousands of ncRNAs were identified in the eukaryotic transcriptome (**Bu et al.** (2011); **Khalil et al.** (2009)). Usually, ncRNAs are divided into two groups according to their length: short ncRNAs (e.g. microRNAs (miRNAs)), consisting of less than 200 nucleotides, and long non-coding RNAs (lncRNAs), sized from 200 nucleotides up to 100 kb (**Mattick** (2001)).
miRNA is the best known class of small RNAs. They are endogenous non-coding regulatory RNAs (17–25 nucleotides), which play important roles in post-transcriptional gene regulation. miRNAs bind the 3'-untranslated regions (3'-UTR) of various target mRNAs leading to direct mRNA degradation or translational repression.
They regulate gene expression and contribute to development, differentiation and carcinogenesis. Many studies have shown that over 30% of human genes are regulated by miRNAs and a single miRNA can control over hundreds of mRNAs. The aberrant expression or alteration of miRNAs also contributes to

a range of human pathologies, including cancer (**Lu et al.** (2005)). Moreover, a significant amount of miRNAs has been found in extracellular human body fluids (**Mitchell et al.** (2008); **Hanke et al.** (2010)) and some circulating miRNAs in the blood have been successfully revealed as biomarkers for several diseases including cardiovascular diseases (**Gupta et al.** (2010)) and cancer (**Mitchell et al.** (2008)).
An emerging class of non coding RNAs consists of long non-coding RNAs (**Fatica and Bozzoni** (2013)). They are both nuclear and cytoplasmic. Nuclear lncRNAs function by guiding chromatin modifiers to specific genomic loci (**Rinn and Chang** (2012); **Batista and Chang** (2013); **Guttman and Rinn** (2012); **Khalil et al.** (2009)) while many others have been identified in the cytoplasm (**Batista and Chang** (2013)). These lncRNAs are involved in gene regulation and often show sequence complementarity with transcripts that originate from either the same chromosomal locus or independent loci.
Interesting, the expression of lncRNAs has been quantitatively analysed in several tissues and cell types by RNA-seq experiments, and it was generally found to be more cell type specific than the expression of protein-coding genes (**Rinn and Chang** (2012); **Derrien et al.** (2012); **Guttman and Rinn** (2012); **Mercer et al.** (2008); **Cabili et al.** (2011); **Pauli et al.** (2012)).
One of the last classes recently discovered is the circular RNA (circRNAs) (**Memczak et al.** (2013)). They are an enigmatic class of RNA with unknown function. Numerous circRNAs form by head-to-tail splicing of exons, suggesting previously unrecognized regulatory potential of coding sequences. Recent results (**Memczak et al.** (2013)) have been shown that thousands of well-expressed, stable circRNAs, often have tissue/developmental-stage-specific expression. Moreover, human circRNAs are bound by miRNAs such as the miR-7 showing a potential role of circRNAs as post-transcriptional regulators.

Understanding the complex system derived from the interactions of regulators and possible targets gives a clue on the dynamics and causes of disorders (**Couzin** (2007)). In this direction, platforms to visualize networks such as Cytoscape (**Shannon et al.** (2003)) together with tools to visualize the topology of regulatory networks are increasing their influence in science.
miRScape (**Ferro et al.** (2009)) is one of the first Cytoscape plugin visualizing protein-protein interaction networks annotated with miRNAs. It uses a web knowledge base (**Laganà et al.** (2009)) to infer associations between genes and phenotypes though miRNAs. CyTargetLinker (**Kutmon et al.** (2013)) is a recent Cytoscape app which builds biological networks annotated with miRNAs, transcription factors and drugs.
Several tools are designed to analyze the regulatory effect of miRNAs in protein coding genes and to export the results in a Cytoscape network format. For example, Magia (**Sales et al.** (2010)) is a web tool to statistically analyze miRNA and gene expressions. TSmir is a web based tool to browse regulatory network of tissue-specific miRNAs with trascriptor factors. mirConnX **Huang et al.** (2011) extend a given network of genes, trascriptor factors and miRNAs with further TF and miRNA-gene intersections inferred by user expression data. miRTrail **Laczny et al.** (2012) statically analyzes the role of miRNAs and genes deregulated in a disease by using an extensive miRNA-gene networks and expression data.

2 CONSTRUCTION AND CONTENT

2.1 DATA SOURCE

In the following sections we describe data sources considered in ncRNAdb.

2.1.1 HGNC The HUGO Gene Nomenclature Committee (HGNC) is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication (http://www.genenames.org/) (**Gray et al.** (2012)).
It contains 19060 protein coding genes, 5714 non-coding RNAs (of those 2546 long non-coding RNAs and 1879 microRNAs), 598 phenotypes and 12621 pseudogenes.
We used HGNC as official database of approved names and alias for each entry in ncRNAdb.

2.1.2 lncRNAdb lncRNAdb provides a central repository of known long non-coding RNAs (lncRNAs) in eukaryotic cells (including those derived from viruses), their aliases and published characteristics (**Amaral et al.** (2011)). lncRNAdb is available online at http://www.lncrnadb.org.
The published version of the database contains over 150 lncRNAs identified from the literature in around 60 different species.

2.1.3 circ2Traits Circ2Traits (http://gyanxet-beta.com/circdb/) is a comprehensive database for circular RNA potentially associated with disease and traits (**Ghosal et al.** (2013)).
Circular RNAs, formed by covalent linkage of the ends of a single RNA molecule, are newly discovered RNAs that sponge miRNAs to block their function (**Memczak et al.** (2013)).
Circ2Traits uses the circular RNA dataset from Memczak et al.(**Memczak et al.** (2013)). This dataset consists of 1953 predicted human circular RNAs along with their genomic coordinates, annotation, and predicted miRNA seed matches. The disease related miRNA data is taken from miR2disease (**Jiang et al.** (2009)). From Starbase (**Yang et al.** (2011)) the authors collected the miRNA-mRNA interaction data predicted by miRanda (**Betel et al.** (2008)), TargetScan (**Lewis et al.** (2005)), PiTA (**Kertesz et al.** (2007)), PicTar (**Krek et al.** (2005)), and RNA22 (**Loher and Rigoutsos** (2012)).
Moreover, dataset for predicted miRNA and long non-coding RNA interaction pairs is collected from the miRCode database (**Jeggari et al.** (2012)).

2.1.4 HMDD The Human microRNA Disease Database (HMDD) (**Li et al.** (2013)) is a database of curated experiment-supported evidence for human microRNAs (miRNAs) and disease associations (http://www.cuilab.cn/hmdd).
Currently, HMDD collected 10368 entries that include 572 miRNA genes, 378 diseases from 3511 papers. The database contains detailed and comprehensive annotations of human miRNA-disease association data, including miRNA-disease association data from the evidence of genetics, epigenetics, circulating miRNAs, and miRNA-target interactions.

2.1.5 lncRNADisease database The LncRNADisease database (http://210.73.221.6/lncrnadisease) (**Chen et al.** (2013)) is a resource for the experimentally supported lncRNA-disease association data and a platform that integrates tools for predicting novel lncRNA-disease associatons.
Moreover, LncRNADisease contains lncRNA interactions in various levels, including proteins, RNAs, miRNAs, and DNA.
The current version of the database has more than 1000 lncRNA-disease entries and 475 lncRNA interaction entries, including 321 lncRNAs and 221 diseases from about 500 publications. LncRNADisease also provides the predicted associated diseases of 1564 human lncRNAs.

2.1.6 miRandola miRandola (http://atlas.dmi.unict.it/mirandola/) (**Russo et al.** (2012); **Russo et al.** (2014)) is the first manually curated database of extracellular circulating miRNAs. It is a comprehensive classification of different extracellular miRNA types and a collection of non-invasive biomarkers for several diseases (e.g. cancer and cardiovascular disease).
The last updated version of the database contains 139 papers, 2366 entries, 599 unique mature miRNAs and several tools for cellular and extracellular miRNA analysis.

2.1.7 NONCODE NONCODE (http://www.noncode.org/) is a database of all kinds of non-coding RNAs (except tRNAs and rRNAs), in particular it contains 210831 lncRNAs for several species (**Bu et al.** (2011)).
NONCODE also provides an ID conversion tool from RefSeq or Ensembl ID to NONCODE ID and a service of lncRNA identification.

2.1.8 NPInter NPInter (http://www.bioinfo.org/NPInter/) (**Wu et al.** (2006)) reports functional interactions between non-coding RNAs (except tRNAs and rRNAs) and biomolecules (proteins, RNAs and DNA) which are experimentally verified.
The authors collected primarily physical interactions, although several interactions of other forms also are included.
Interactions are manually collected from publication, followed by an annotation process against known databases including NONCODE (**Bu et al.** (2011)), miRBase (the miRNA registry) (**Kozomara and Griffiths-Jones** (2013)) and UniProt (the database of proteins) (**Consortium et al.** (2013)).
The second version of NPInter contains 201107 interactions of ncRNA with other biomolecules from 18 organisms.

2.2 DATA SCHEMA

Inside public databases, biological entities (e.g. genes) are catalogued via common used nomenclatures, they can be human readable names or alphanumeric identifiers (that include accession numbers), with the intent to assign a keyword that can identify the specific entity. For example, genes are classified by their names (usually, a summary description of their function), by their symbols (short abbreviations) or database-specific identifiers. The *breast cancer 1* gene can be identified by its assigned symbols BRCA1, BRCC1 and PPP1R53, or by its specific identifier among gene databases like HGNG:1100, Entrez Gene 672 and UCSC uc002ict.3. Unfortunately, some of these nomenclatures are not disambiguous, thus distinct biological entities can be identified by the same nomenclature. The gene example is a lucky one, like the disease nomenclature, since, in the last years, researches have made an effort for their disambiguation, like the HUGO Gene Nomenclature Committee. Non-coding RNAs, as we present here, are a relatively recent discovery and there is not yet a comprehensive database including all of them and where the researcher can find an unambiguous nomenclature systems. Moreover, the non-coding RNA knowledge is spread among several databases where new discovered entities are named with internal identifiers and poor mapping schemas between such databases exist. The largest collection of ncRNA is NONCODE v4 but the mapping, from internal to external ID systems, that it provides is still not extensive also because most the reported ncRNAs are uniquely to NONCODE and they are not reported in any other database.
For our purpose, we avoided the use of an our internal nomenclature system but we introduced a generic resource identifier system. Each identifier is composed by three parts, also called levels: the entity type, the data source and the external name, respectively.

$$entitytype : datasource : alias \tag{1}$$

The entity type describes the biological classification of the element, while the data source describes the external data resource from which the information come from. An entity type can be NCRNA, DISEASE, GENE or RNA (which does not include ncRNA). A data source, or resource, names is composed by the name of the external resource plus its version (e.g. HMDD_2). We also consider two additional data source identifiers, GENERIC and UNKNOWN, for dealing with generic names (e.g. the symbol BRCA1 is a generic used one that can not be related to any specific resource) and unknown (to us) resources, respectively. The resource identifiers is also used to store the origin of a relation, namely the databases from where it has been extracted. Table **??** shows the complete set of resource identifiers used in our system. Moreover, the database engine (OrientDB) used for out system labels each item with a unique system-scope identifier, called ORID, and every record (entity or relation) can be retrieved by its ORID. Thus, our record identify format allows to retrieve and entity by its ORID in place of a three level identifier.

A system representing a set of biological entities and their relations (physical interactions, functional relations, etc...) can be naturally modelled within a network, or formally described as a graph, a mathematical object composed by vertices (entities) and edges (relations). Classical relation database management systems (RDMS) are widely used to store biological data. Such systems are based on the relational model which is designed to represent an entity-relation system, and often the SQL (Structure

Query Language) language is used to provide a querying interface to data. However, newly models have been proposed to solve performance issues or to provide additional feature in specific applications. Some of them are grouped under the name of NoSQL (Not only SQL) databases and they are increasingly used in big data and web applications, since they can provide schema-less representation for non-structured data. Nowadays, there a re plenty of such systems, each one with specific features and data model for data representation. They can be summary classified by four main data models, even if some of them include more then one model: column model where data are represented by tuples, document-oriented databases for storing, retrieving, and managing document-oriented information, also known as semi-structured data, key-value model where data are stored as a collection of key-value pairs stored using associative arrays, maps, symbol tables of dictionaries, and finally graph databases where data are modelled using a graph structure. Moreover, they often implements the object-oriented model since they are used as persistence layer in developing software using object-oriented languages, thus they model concepts like class, inheritance and polymorphism. Functionality, complexity, flexibility, scalability and performance depend on the suitability of the implemented data models to specific applications. Usually, RDMS systems store data into tables and provide a single table index (based on the key columns of the table) for querying data in an efficient way. Instead, a pecuniary feature of NoSQL systems is that several indexes can be built on the same data container, proving different access point to the same data.

Among the several NoSQL databases, OrientDB has been chosen as persistence layer of our system. It implements a graph model and an object-oriented model, on top of a document model. Moreover, it offers an SQL interface in addition to several language specific interfaces. It is developed in Java and provides native Java API (Application Programming Interface) for accessing the database, which is suitable for developing Cytoscape applications. Figure 1 shows the schema of our proposed systems. We preferred to keep separated the concepts of biological entity and its related aliases. The abstract class *BioEntity* represents biological entities and it specializes in the four biological entity types ncRNA, Gene, Disease and RNA. Aliases are represented by the abstract class *Alias* which is specialized in four different sub-classes related to the four biological entity types. Each alias stores the coming resource and the internal ID (in such resource), thus the complete three levels identifier can be reconstruct combining those two fields with the specific alias sub-class where they are stored. In a graph model, elements on both classes (and sub-classes) represent vertices of the graph having distinct properties. The naming of a biological entity by an alias is represented by adding an edge between the corresponding graph vertices. Aliases serve as access points to the biological entity collection. Two different indexes are built on top of each alias sub-class. The first one is a composite key dictionary working on the two fields, resource and ID, to provide an unique identification of an external reference, and it implements the fully three levels alias identification. The second one is a single field not-unique map to efficiently query the set of IDs of an alias sub-class without combining them to the external resource identification. This index results useful when the origin (the resource where it comes from) of an ID is not known, thus the system outputs all the aliases having such ID regardless the external resource system that produced them. In this way, we can implement a two levels (ambiguous) identification system based on the biological type and a generic name. A further not-unique index works on the entire *Alias* class hierarchy to implement a single level identification when neither the biological entity type ans the external resource are known. This index was introduced because sometimes some of the imported databases do not specify the biological entity type and we were not able to catalogue some elements because no matches in other databases where found or because they were ambiguous. Interactions are modelled as graph edges between two elements of the class *BioEntity*. For each interaction ad additional set of information is stored, when the imported databases included them. It include a PubMed ID of the referencing article plus the phrase supporting the relation, the database from where the interaction was extracted and the interaction level. The interaction level is composed by a pair of elements within $\{RNA, DNA, Protein, TF\}$, on NA (when it is not specified). For example, the pair $RNA - TF$ specifies that a ncRNA is interacting with the transcription factor of a gene, while the pair $RNA - DNA$ specifies that the ncRNA is interacting with the coding genomic region of the gene. If the same relation is stored in two (or more) distinct data sources, than two (or more) interaction edges are present inside our system.

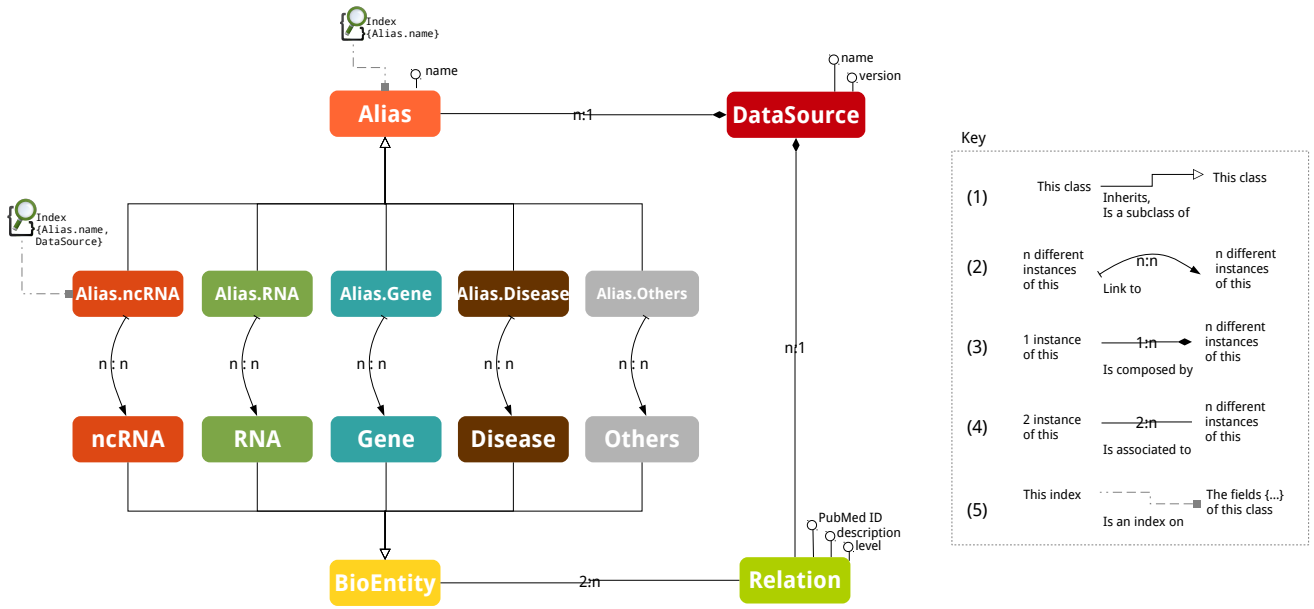

Figure 1. The database schema.

2.3 DATA IMPORT

In this section, we describe the import phase from the external databases listed in section **??** to our system. The main issue of such operation was the alias ambiguation and the entity type miss-knowledge. The first one is described in section 2.2 and it has been resolved by a merging procedure. Formally, given three aliases A,B and C such that two different mappings $A \mapsto B$ and $B \mapsto C$, possibly coming from different data sources, have been found. Then the three aliases are considered as alternative nomenclatures to the same biological entity. This procedure does not prevents that the first two aliases are referred to a different biological entities of the other two, for example when a gene name is used to identify its transcript, but reduces the number of mismatching where querying the system. The miss-knowledge issue wad due to the fact that the imported interaction data were in the form ncRNA-(something else), and the biological type of the second interaction actors was not specified, neither the indication of the external nomenclature system. Thus, we were not able to fully classify (with a complete three level identification) the entity, and it has been catalogued as $UNKNOWN$ type. Some of such miss-classification were resolved by applying a procedure after the import step, when all the data from the external sources were imported and more information was available, by trying to reclassify the $UNKNOWN$ type entities.

Below, a list of the importing data source and the type of information, extracted from them, is given. The order is the same in which with extract and add up information from a source to our collection. For all the importing source, we considered only those information related to the species Homo sapiens.
From HGNC (the database of human gene names) we import a list of non-coding RNA, protein-coding genes, pseudogenes and phenotypes (considered as diseases) and their approved aliases in a series of well-established databases.
From lncRNAdb we import a list of non-coding RNA and their common names (aliases, symbols). This source includes also some briefly description about expression, function, conservation and misc that we do not take into account.
From circ2traits, we import a set of interacting lncRNA, circRNA and messanger RNA together with the disease, relation and article where the interaction is reported.
From HMDD, we import a list of disease names together with a set of genes that interact with a ncRNA.

In this database, interactions are listed in the forms ncRNA-disease or ncRNA-gene-disease, and the referencing article together the support sentence are reported. The ncRNA-gene-disease multi-relations are split into distinct relation of type ncRNA-gene and ncRNA-disease.
From lncRNAdisease, we import a list of lncRNA (and their symbol aliases) and diseases and their (co-)relation supported by a pubmed reference, a dysfunction type description and a support sentence. Moreover, from this database, we import also a set of entities interactions supported by the article, dysfunction type, support sentence and interaction level. In this case, since the entity type the actors of the interaction is not provided, but only the entity level (i.e. RNA-protein) at first we query our system (filled with the information extracted until this step) searching for a corresponding entity to the given alias, and if no entry is found then the new alias is stored and connected to a new entity of unknown type.
From the Mirandola database, we import a set of miRNA (and their aliases) and their related disease together with the supporting article and sentence.
From NONCODE, we import a list of non-coding RNA aliases and a mapping from the NONCODE v4 DI system to other ID systems. Moreover, we import also a set of associated GO terms related to each listed ncRNA.
From NPInter, we import a set of ncRNA and their interactions with other entities supported by the referencing article, support sentence and interaction level. Similar to the lncRNAdisease case, the entity type of the second actor of the interaction is not specified thus at first we search for an entry in our database and if the entity alias is still not reported we insert e new unknown-type entity and its alias.

At the end of the import procedure, $853,543$ three-level identifiers (alias) were created, a total of $222,970$ biological entities were imported, and $889,675$ edges between *Alias* and *BioEntity* classes were created. Moreover, $238,524$ entity relation were extracted. Table **??** shows the total number of imported biological entities, sorted by they biological type, and how many of them are actually involved in a relation.

3 UTILITY AND DISCUSSION

As said before, OrientDB is supported by plenty of language connectors, beside the native Java API. One a server instance is up, the final user can connect to it by a series of programming language binding, or by using the OrientDB SQL console. It also implements technology standard like HTTP REST/JSON, TinkerPop Blueprints (for graph computing) and JDO (Java Data Object for object persistence).
However, we propose three alternative interfaces available to final users for querying our public database instance. A CytoScape (version 3) plug-in for importing data in a network environment, plus a web interface and a command-line interface for raw resource queries. The CytoScape plug-in and the command-line applications can be downloaded from the ncRNA-DB website, while the web interface is part of it. A complete schema of the proposed system, from the import phase to the user-end interfaces, is shown in figure 2.

3.1 CYTOSCAPE INTERFACE

The CytoScape interface is composed by three main functions, integrated in a single CytoScape plug-in, for expanding an input network with ncRNA-DB data, searching for specific biological entities, or importing the whole ncRNA-DB in a CytoScape network environment. Figure 3 shows the graphical user interfaces of such applications, that can be accessed through the *ncRNA-DB* menu item, once the plug-in have been installed.

The *Expand Network* functionality takes in input a user built network and expand (add to) it ncRNA-DB entities and relations. It is required that the input network table must have two columns for specifying the biological entity type of an element and for listing its set of known aliases in their ncRNA-DB

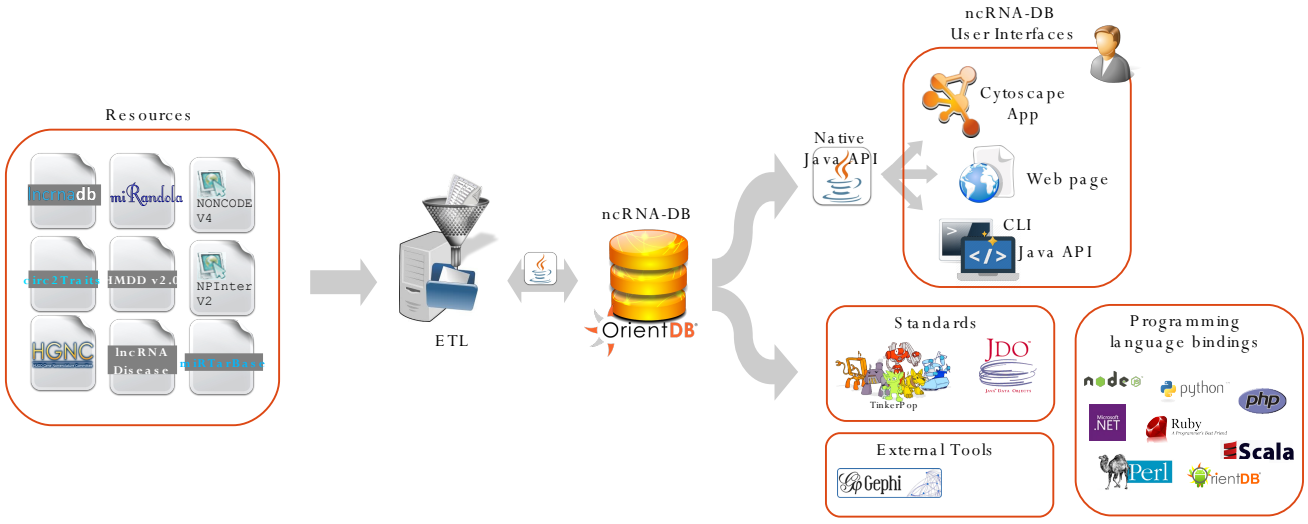

Figure 2. Complete system architecture and procedures.

identify format. The entity type can be one of $NCRNA$, RNA, $GENE$ or $DISEASE$ for fully qualified resource retrieving, or it can be a blank field, in this case the application will search for al the biological entity types. If the key $GROUP$ is used as entity type, then the application associates all the matching aliases to the node. The user can decide whenever some biological entity types must be excluded from the retrieving by unflagging the corresponding check-boxes. Once those two columns have been specified, the application retrieves the matching biological entities, maps them to the CytoScape nodes and adds to the selected network the set of found relations between the retrieved nodes. If the aliases set of a CytoScape node matches with more then one ncRNA-DB entity, then the node is mapped to all of such entities. This behaviour allows the user to specify abstract nodes representing and entity group, for example, one can specify a group of aliases of different biological type and compact them in a single entity node. It also helps to eliminate ambiguity at this level of data representation.
If the *Include neighbours* check-box is flagged, then the application retrieves all the ncRNA-DB neighbours of the matching entities and adds them to the selected CytoScape network, as well as relations between them an the other retrieved entities.

The second function, *List entities*, allows the user to search for particular biological entities by specifying their aliases (fully or incomplete ncRNA-DB identifiers). Aliases can be put in the text form of the application or loaded from file. If aliases are specified by the text form, then a single CytoScape node is created, instead, if they are specified from file then a node is created for each file row. The file format expects a series of rows, each of which can contain one or more aliases. The application retrieves the matching entities and the user can select a list of them to be imported in the selected CytoScape network. For each entity the complete list of corresponding aliases, those taken in input plus those retrieved from the database, and its biological type are shown. Figure 4 showns the querying page of the proposed web interface.

Finally, the *Import DB* application allows to import the whole ncRNA-DB database, or just a part of it by specifying a series of discriminating rulers. The user can select which of the biological entity type must be imported and which kind of relations. The retrieved elements are imported in a new Cytoscape network having the user specified name. We suggest to use this functionality properly, since hundreds of

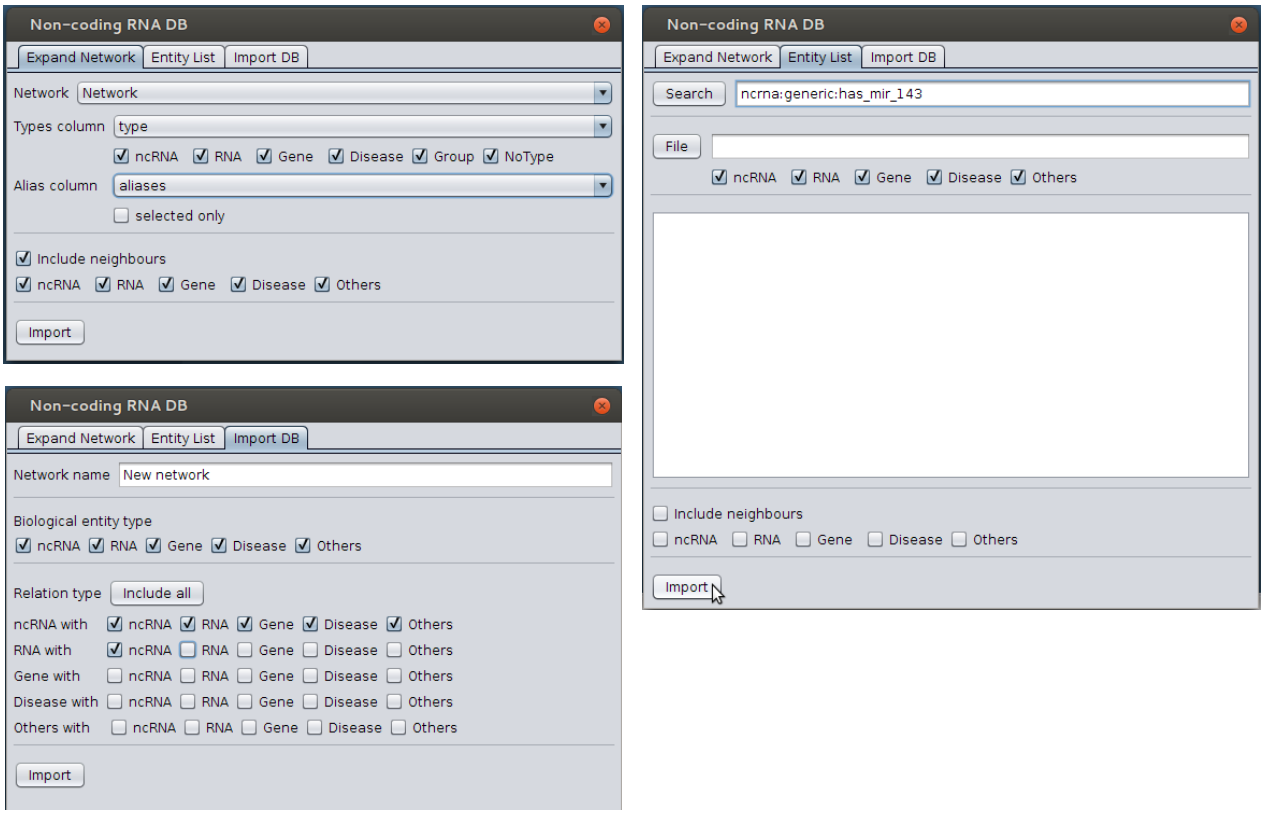

Figure 3. The Cytoscape plug-in.

Figure 4. The main page of the web interface.

thousands of items, both entities and relations, are stored in out databases, and the cause performance of minimal resource requirement issues when imported all together in the CytoScape environment.

3.2 WEB INTERFACE

We developed a web interface for querying our databases. On the main page, the user specify a set of entity identifiers (as described in section 2.2) or an internal ncRNA-DB ORID, then the matching entities are printed on the result panel. For each entity, the set of its alias is shown, as well as its internal ncRNA-DB ORID, and a link for viewing the set of interacting entities.

3.3 COMMAND-LINE INTERFACE

We developed a command-line interface to our database for entity searching and relation retrieval. It is released as a Java package to be platform independent and it does not require any external dependency. It provides to different commands for accessing the database data. The *search* command takes as input a list of entity coordinates and returns the matching biological entities stored in the databases. This command helps to verify if an identifiers is included in the database and to retrieve all its alternative nomenclatures which were extracted during our data import phase. Moreover, it can be used to understand how a nomenclature has been catalogued. For example, a one level coordinate, e.g. a gene alias, search

helps to recognize different data sources using the same identifier or whenever the alias is used for distinct biological entity types. The second command, *relations*, receives as input a list of entity coordinates and outputs the relations between them stored in the database and their support information. If more than one biological entities match the same input coordinate, then they are all included in the output.

4 CONCLUSION

DISCLOSURE/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.

AUTHOR CONTRIBUTIONS

The statement about the authors and contributors can be up to several sentences long, describing the tasks of individual authors referred to by their initials and should be included at the end of the manuscript before the References section.

ACKNOWLEDGEMENT

Funding: Francesco Russo has been supported by a fellowship sposored by 'Progetto Istituto Toscano Tumori Grant 2012 Prot.A00GRT'.

REFERENCES

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2011), lncRNAdb: a reference database for long noncoding RNAs, *Nucleic acids research*, 39, suppl 1, D146–D151

Batista, P. J. and Chang, H. Y. (2013), Long noncoding RNAs: cellular address codes in development and disease, *Cell*, 152, 6, 1298–1307

Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008), The microRNA.org resource: targets and expression, *Nucleic acids research*, 36, suppl 1, D149–D153

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., et al. (2011), NONCODE v3. 0: integrative annotation of long noncoding RNAs, *Nucleic acids research*, gkr1175

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011), Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, *Genes & development*, 25, 18, 1915–1927

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2013), LncRNADisease: a database for long-non-coding RNA-associated diseases, *Nucleic acids research*, 41, D1, D983–D986

Consortium, E. P. et al. (2012), An integrated encyclopedia of DNA elements in the human genome, *Nature*, 489, 7414, 57–74

Consortium, U. et al. (2013), Update on activities at the Universal Protein Resource (UniProt) in 2013, *Nucleic acids research*, 41, D1, D43–D47

Couzin, J. (2007), Erasing microRNAs reveals their powerful punch, *Science*, 316, 5824

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012), The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression, *Genome research*, 22, 9, 1775–1789

Fatica, A. and Bozzoni, I. (2013), Long non-coding RNAs: new players in cell differentiation and development, *Nature Reviews Genetics*

Ferro, A., Giugno, R., Laganà, A., Mongiovì, M., Pigola, G., Pulvirenti, A., et al. (2009), miRScape: A Cytoscape Plugin to Annotate Biological Networks with microRNAs, in Network Tools and Applications in Biology (NETTAB), Focused on Technologies, Tools and Applications for Collaborative and Social Bioinformatics Research and Development

Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013), Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits, *Frontiers in genetics*, 4

Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., and Bruford, E. A. (2012), Genenames.org: the HGNC resources in 2013, *Nucleic acids research*, gks1066

Gupta, S. K., Bang, C., and Thum, T. (2010), Circulating microRNAs as biomarkers and potential paracrine mediators of cardiovascular disease, *Circulation: Cardiovascular Genetics*, 3, 5, 484–488

Guttman, M. and Rinn, J. L. (2012), Modular regulatory principles of large non-coding RNAs, *Nature*, 482, 7385, 339–346

Hanke, M., Hoefig, K., Merz, H., Feller, A. C., Kausch, I., Jocham, D., et al. (2010), A robust methodology to study urine microRNA as tumor marker: microRNA-126 and microRNA-182 are related to urinary bladder cancer, in Urologic Oncology: Seminars and Original Investigations, volume 28 (Elsevier), volume 28, 655–661

Huang, G., Athanassiou, C., and Benos, P. (2011), mirConnX: condition-specific mRNA-microRNA network integrator, *Nucleic Acids Research (Web server issue)*, 39

Jeggari, A., Marks, D. S., and Larsson, E. (2012), miRcode: a map of putative microRNA target sites in the long non-coding transcriptome, *Bioinformatics*, 28, 15, 2062–2063

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009), miR2Disease: a manually curated database for microRNA deregulation in human disease, *Nucleic acids research*, 37, suppl 1, D98–D104

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007), The role of site accessibility in microRNA target recognition, *Nature genetics*, 39, 10, 1278–1284

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., et al. (2009), Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, *Proceedings of the National Academy of Sciences*, 106, 28, 11667–11672

Kozomara, A. and Griffiths-Jones, S. (2013), miRBase: annotating high confidence microRNAs using deep sequencing data, *Nucleic acids research*, gkt1181

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., et al. (2005), Combinatorial microRNA target predictions, *Nature genetics*, 37, 5, 495–500

Kutmon, M., Kelde, T., Mandaviya, P., Evelo, C. T., and Coort, S. L. (2013), CyTargetLinker: A Cytoscape App to Integrate Regulatory Interactions in Network Analysis, *PLoS One*

Laczny, C., Leidinger, P., Haas, J., Ludwig, N., Backes, C., Gerasch, A., et al. (2012), miRTrail - a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases, *BMC Bioinformatics*, 13, 36

Laganà, A., Forte, S., Giudice, A., Arena, M., Puglisi, P., Giugno, R., et al. (2009), miR: a miRNA knowledge base, *DATABASE*

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005), Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *cell*, 120, 1, 15–20

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2013), HMDD v2. 0: a database for experimentally supported human microRNA and disease associations, *Nucleic acids research*, gkt1023

Loher, P. and Rigoutsos, I. (2012), Interactive exploration of RNA22 microRNA target predictions, *Bioinformatics*, 28, 24, 3322–3323

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005), MicroRNA expression profiles classify human cancers, *nature*, 435, 7043, 834–838

Mattick, J. S. (2001), Non-coding RNAs: the architects of eukaryotic complexity, *EMBO reports*, 2, 11, 986–991

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013), Circular RNAs are a large class of animal RNAs with regulatory potency, *Nature*, 495, 7441, 333–338

Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. (2008), Specific expression of long noncoding RNAs in the mouse brain, *Proceedings of the National Academy of Sciences*, 105, 2, 716–721

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., et al. (2008), Circulating microRNAs as stable blood-based markers for cancer detection, *Proceedings of the National Academy of Sciences*, 105, 30, 10513–10518

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012), Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis, *Genome Research*, 22, 3, 577–591

Rinn, J. L. and Chang, H. Y. (2012), Genome regulation by long noncoding RNAs, *Annual review of biochemistry*, 81

Russo, F., Di Bella, S., Bonnici, V., Laganà, A., Rainaldi, G., Pellegrini, M., et al. (2014), A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs, *BMC Genomics*, 15, 3, 1–7

Russo, F., Di Bella, S., Nigita, G., Macca, V., Lagana, A., Giugno, R., et al. (2012), miRandola: extracellular circulating microRNAs database, *PLoS One*, 7, 10, e47786

Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., and Romualdi, C. (2010), MAGIA, a web-based tool for miRNA and Genes Integrated Analysis, *Nucleic Acids Res*

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., et al. (2003), Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Research*, 13, 11, 2498–504

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., et al. (2006), NPInter: the noncoding rnas and protein related biomacromolecules interaction database, *Nucleic acids research*, 34, suppl 1, D150–D152

Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., and Qu, L.-H. (2011), starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data, *Nucleic acids research*, 39, suppl 1, D202–D209

FIGURES

Figure 1. Enter the caption for your figure here. Repeat as necessary for each of your figures.