Universidad de Murcia
Máster en Bioinformática

LncRNA-fibrotic diseases association dataset

Silvia García Juan

All data in /home/alumno15/entrega_semantica
December 12th, 2022

Contents Index

Figures Index

Tables Index

1. Background

The RDF dataset created for this project is based on the Fibrotic Disease-associated RNAome database (FDRdb) (http://www.medsysbio.org/FDRdb), released for the first time on March 15th, 2021, and whose last update is from April 6th, 2022. This last version includes 1950 associations among 912 RNAs and 92 fibrotic diseases in 8 species. All this data has been collected from 1127 published literature.

The FDRdb is available for free and it conforms a valuable resource for researchers to explore the mechanisms of RNA dysregulation in organ fibrosis.

Fibrotic diseases are common pathologies encompassing a wide spectrum of clinical entities which have in common severe tissue injury or dysregulation of wound-healing repair. Fibrosis is defined as the abnormal deposition of the extracellular matrix and involves several organs such as the heart, kidney lung, liver, and skin(Figure 16). Fibrotic diseases can lead to organ dysfunction and death.

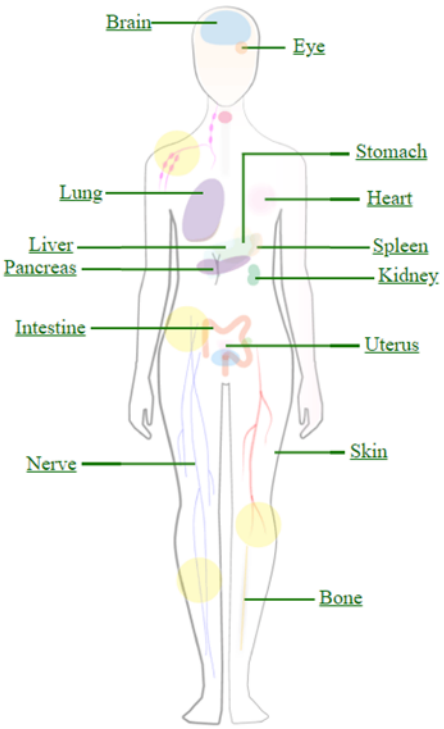

Figure 1: Body organs documented in FDRdb where fibrotic diseases appear.

Due to the complexity of this disease and its multiple pathogenic factors, the primary challenge of research is focused on the mechanism underlying fibrosis. The RNAome involved in fbrotic diseases includes two categories: coding (RNA) and non-coding RNA (ncRNA) and their up and down-regulation are correlated with the development of the illness.

With the aim of improving the usability of the data set and make it more accessible, integrated, and easy to manage, part of the data from FDRdb has been developed into a semantic dataset. This has been possible carrying of the FAIRification of the data, by adhering to the guiding principles of

Linked Data and using RDF as the representation language.

2. Creation of the RDF dataset

2.1. Filtering of the dataset file

From the FDRdb dataset, only the "LncRNA-fibrotic diseases association data" from *Mus musculus* species was downloaded. The file was a CSV with 131 rows and 18 columns. The list of the 13 columns selected from this file was as follows:

- **Tissue**. Name of the tissue in which the disease appears.
- **Disease**. Name of the fibrotic disease or the disease provoked by the fibrosis.
- **Species**. Species in which the disease appears. In this case all the column was *Mus musculus*.
- **Species**. Species in which the disease appears. In this case all the column was *Mus musculus*.
- **Symbol**. Symbol of the RNA.
- **ID**. The RNA ID in the Ensembl database.
- **Type**. RNA type. In this case all this column is lncRNA.
- **Expression**. If the ARN is up-regulated or down-regulated in the development of the disease.
- **RNA target**. Target of the RNA molecule.
- **Pathway**. Pathway in which the RNA participates.
- **Function**. Function of the RNA in the process.
- **PMID**. The publication ID in the PubMed database in which the ARN-fibrotic disease association appears
- **Title**. Title of the publication in which the ARN-fibrotic disease association appears.
- **Year**. Year of publication of the study.

All the 131 rows from the CSV file where selected.

2.2. Creation of the triplets

For the creation of the triples, the Turtle syntax was used, which would be converted to RDF syntax in the following step.

A templated file with the desired triplets was created manually and focusing on the first row in the CSV file. The URIs from Classes and Properties were collected from the EMBL-EBI Ontology Look-up Services (OLS) (https://www.ebi.ac.uk/ols/index). 10 of the ontologies were chosen:

- RDF
- RDF Schema
- OWL 2 Web Ontology Language
- Dublin Core (DCMI)
- NCIT Ontology
- Semanticscience Integrated Ontology (SIO)
- Biological Collections Ontology (BCO)
- BioAssay Ontology (BAO)
- Ensembl Glossary
- Common Coordinate Framework (CCF) Ontology

The URI from the graph is <http://lncrna.com/graph/lncrna_mus> and the entities from the graph have as identifier < http://lncrna.com/resources/> followed by the id of the entity.

The triplets created have the following shape:

2.2.1. Publications

Each publication has 4 properties, so resulted in 4 triplets for publication. The URIs from the subjects were formed by <http://lncrna.com/resources/PUB_(ID) >, being the ID the row from the CSV where the publication appears.

Table 1: Publications properties.

Property	**Property URI**
1.title	<http://purl.org/dc/elements/1.1/title>
2.date	<http://purl.org/dc/elements/1.1/date>
3.sameAs	<http://www.w3.org/2002/07/owl#sameAs>
4.class	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

Table 2: Publications objects.

Object	**Object URI**
1.Publication title	Literal without URI corresponding to the column "Title"
2.Year	Literal without URI corresponding to the column "Year"
3.PubMed Entry	<https://pubmed.ncbi.nlm.nih.gov/(ID corresponding to the column "PMID")
4.NCIThesaurus "publication" term entry	<purl.obolibrary.org/obo/NCIT_C47902>

2.2.2. RNAs

Each RNA has 11 properties, so resulted in 11 triplets for RNA. The URIs from the subjects were formed by <http://lncrna.com/resources/ARN_(ID) >, being the ID the row from the CSV where the publication appears. For space reasons I have created two different tables for properties and objects, whose numbers match when they are in the same triplet.

Table 3: RNAs properties.

Property	**Property URI**
1.results in	<http://semanticscience.org/resource/SIO_ 001156 >
2.sameAs	<http://www.w3.org/2002/07/owl#sameAs>
3.type	<http://purl.org/dc/elements/1.1/>
4.has target	<http://semanticscience.org/resource/SIO_000291>
5.Pathway	<http://rs.tdwg.org/dwc/terms/pathway>
6.has function	<http://semanticscience.org/resource/SIO_000225>
7.label	<http://www.w3.org/2000/01/rdf-schema#>
8.organism	<http://purl.uniprot.org/core/organism>
9.class	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
10.Name Published In	<http://rs.tdwg.org/dwc/terms/namePublishedIn>
11.has associated disease	<http://www.bioassayontology.org/bao#BAO_0002848>

Table 4: RNAs objects.

Object	Object URI
1.Expression of the RNA	Literal without URI (up-regulation or down-regulation) corresponding to the column "Expression"
2.Ensembl ARN entry	<http://ensembl.org/gene/(ID corresponding to the column "ID")>
3.NCIThesaurus "transcript" term entry	<http://purl.obolibrary.org/obo/NCIT_C88924>
4.ARN Target	Literal without URI corresponding to the column "Target"
5.Pathway	Literal without URI corresponding to the column "Pathway"
6.Function	Literal without URI corresponding to the column "Function"
7.Symbol	Literal without URI corresponding to the column "symbol"
8.Uniprot Taxonomy Mus musculus entry	<https://www.uniprot.org/taxonomy/39442>
9.NCIThesaurus "transcript" term entry	<http://purl.obolibrary.org/obo/NCIT_C1936>
10.Publication in which it appears	<http://lncrna.com/resources/PUB_(ID)>
11.Disease which it has associated	<http://lncrna.com/resources/DISEASE_(ID)>

2.2.3. Organism

Each organism taxon forms 1 extra triplet because it has its own label.

Table 5: Organisms subjects.

Subject	Subject URI
1.Uniprot Taxonomy Mus musculus entry	https://www.uniprot.org/taxonomy/39442

Table 6: Organisms properties.

Property	Property URI
1.label	<http://www.w3.org/2000/01/rdf-schema#>

Table 7: Organisms objects.

Object	Object URI
1.Organism	Literal without URI corresponding to the column "Species"

2.2.4. Diseases

Each Disease has 4 properties, so resulted in 4 triplets for Disease. The URIs from the subjects were formed by <http://lncrna.com/resources/DISEASE_(ID) >, being the ID the row from the CSV where the publication appears.

Table 8: Diseases properties.

Property	**Property URI**
1.label	<http://www.w3.org/2000/01/rdf-schema#>
2.tissue provider name	<http://purl.org/ccf/tissue_provider_name>
3.class	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
4.is associated disease of	<http://www.bioassayontology.org/bao#BAO_0002849>

Table 9: Diseases objects.

Object	**Object URI**
1.Disease	Literal without URI corresponding to the column "Disease"
2.Tissue	Literal without URI corresponding to the column "Tissue"
3.NCIThesaurus "disease or disorder" term entry	http://purl.obolibrary.org/obo/NCIT_C2991
4.ARN which is an associated disease of	<http://lncrna.com/resources/ARN_(ID)>

Once this template with the first-row data was created, it was converted to RDF syntax and validated as a graph. Then, the template was modified to serve as an actual template, substituting the data with its column name.

With the python program csv_to_ttl.ipynb placed in the /home/alumno15/entrega_semantica'folder, the template was used to make triplets from every row in the CSV with the replace function as it can be observed. The output of the program was a TTL file without the prefixes, which were pasted manually to the file. This resulting TTL was converted to RDF syntax and 2490 triplets were obtained.

2.3. Metadata

A file with the graph metadata was created in turtle syntax and the file is *Metadata.ttl*.

3. Upload the RDF database

A new graph with the name *lncRNA_mus* was created in the Blazegraph graph database. Then the RDF file obtained as output in the previous step was uploaded as a n-triples file, and the file is

lncrna_mus.nq with the update function in the web page. The metadata was also uploaded with the same function.

4. SPARQL Queries

4.1. Blazegraph Queries

4.1.1. Query 1

Get all the upregulated ARNs.

```
PREFIX lncrna: <http://lncrna.com/resources/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX SIO: <http://semanticscience.org/resource/SIO_>

SELECT DISTINCT ?ARN ?regulation
WHERE {
 ?instance rdfs:label ?ARN ;
 SIO:001156 ?regulation ;
 SIO:001156 "up-regulation" .
}
```

Figure 2: Query 1 input.

ARN	regulation
Miat	up-regulation
Malat1	up-regulation
H19	up-regulation
NONMUT021928.2	up-regulation
Chrf	up-regulation
Trp53cor1	up-regulation
Gpr137b-ps	up-regulation
Snhg7	up-regulation
Zfas1	up-regulation
Mhrt	up-regulation
Norad	up-regulation
Gas5	up-regulation
Fendrr	up-regulation
Isg20	up-regulation
Xist	up-regulation
Rmrp	up-regulation
Dlx6os1	up-regulation
Rassf1	up-regulation
Sox2ot	up-regulation
AK081284	up-regulation
Neat1	up-regulation
Arid2	up-regulation
Gm4419	up-regulation
Hottip	up-regulation
Hotair	up-regulation
Tug1	up-regulation
1500026H17Rik	up-regulation
6030443J06Rik	up-regulation

Figure 3: Query 1 output.

Results	Execution Time
43	111ms

Figure 4: Query 1 results.

4.1.2. Query 2

Get the names of every ARN, with at least two functions, alongside their total number of functions and sorted by the number of functions.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX lncrna: <http://lncrna.com/resources/>
PREFIX SIO: <http://semanticscience.org/resource/SIO_>

SELECT ?ARN (COUNT(?ARN) As ?noOfFunctions)
WHERE {
  ?instance rdfs:label ?ARN ;
            SIO:000225 ?function .

  ?instance rdf:type NCIT:C1936 .
}
GROUP BY ?ARN
HAVING (?noOfFunctions > 1)
ORDER BY DESC(?noOfFunctions)
```

Figure 5: Query 2 input.

ARN	noOfFunctions
H19	11
Gas5	8
Malat1	7
Miat	6
Neat1	5
Trp53cor1	4
Xist	3
Mir17hg	2
Hotair	2
Gpr137b-ps	2
Snhg7	2
Sail	2
Zfas1	2
Norad	2
Rmrp	2
Sox2ot	2
Hottip	2
Tug1	2
D630029K05Rik	2
Snhg6	2
Kcnq1ot1	2
Meg3	2

Figure 6: Query 2 output.

Results	Execution Time
22	100ms

Figure 7: Query 2 results.

4.1.3. Query 3

Get each disease alongside the number of lncRNA they are related to and sorted by the number of lncRNA.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX BAO: <http://www.bioassayontology.org/bao#BAO_>

SELECT ?Disease
       (COUNT(?ARN) As ?noOflncRNA)
WHERE {
 ?instance rdfs:label ?Disease ;
           rdf:type NCIT:C2991.

 ?transcript rdf:type   NCIT:C1936 ;
             BAO:0002848 ?instance ;
             rdfs:label ?ARN .
}
GROUP BY ?Disease
ORDER BY DESC(?noOflncRNA)
```

Figure 8: Query 3 input.

Disease	noOflncRNA
diabetic_nephropathy	28
renal_fibrosis	26
liver_cirrhosis	19
endomyocardial_fibrosis	13
myocardial_infarction	7
idiopathic_pulmonary_fibrosis	4
diabetic_cardiomyopathy	4
heart_failure	4
nephritis_interstitia	4
kidney_diseases	3
kidney_failure_chronic	3
pulmonary_fibrosis	2
respiratory_distress_syndrome_acute	2
acute_kidney_injury	2
carcinoma_hepatocellular	1
cardiovascular_diseases	1
atrial_fibrillation	1
pulmonary_hypertension	1
silicosis	1
acute_myocardial_infarction	1
atrial_fibrosis	1
chronic_kidney_disease	1
ventricular_dysfunction	1
colorectal_neoplasms	1

Figure 9: Query 3 output.

4.1.4. Query 4

Get every function with “inhibits” in their name alongside the ARN that carries on this function and year that it was publicated.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX BCO: <http://rs.tdwg.org/dwc/terms/>
PREFIX SIO: <http://semanticscience.org/resource/SIO_>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>

SELECT DISTINCT ?function ?ARN ?year
WHERE {
 ?instance SIO:000225   ?function ;
           #rdf:type NCIT:C1936 ;
           rdfs:label   ?ARN ;
           BCO:namePublishedin ?publication .

  ?publication dc:date ?year .

  filter contains (?function,"inhibits")
}
```

Figure 10: Query 4 input.

function	ARN	year
Rassf1-as1 inhibits the translation of rassf1a to exacerbate cardiac fibrosis in mice.	Rassf1	2019
Knockdown of lncrna gas5 leads to antifibrosis by competitively binding miR-96-5p. which inhibits the expression of FN1.	Gas5	2020

Figure 11: Query 4 output.

Results	Execution Time
2	155ms

Figure 12: Query 4 results.

4.1.5. Query 5

Get the publications of 2020 related with a LncRNA-fibrotic diseases association lncRNA along with the title, year, disease and PubMed link.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX BCO: <http://rs.tdwg.org/dwc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX BAO: <http://www.bioassayontology.org/bao#BAO_>
PREFIX owl: <http://www.w3.org/2002/07/owl#>


SELECT ?title ?year ?disease ?pubmed
WHERE {
?publication dc:title ?title ;
             dc:date ?year ;
             owl:sameAs ?pubmed .
?arn BCO:namePublishedin ?publication ;
     BAO:0002848 ?dis .
?dis rdfs:label ?disease .

  filter contains (?year, "2020")
}
```

Figure 13: Query 5 input.

title	year	disease	pubmed
LncRNA MIAT/miR-133a-3p axis regulates atrial fibrillation and atrial fibrillation-induced myocardial fibrosis	2020	atrial_fibrillation	<https://pubmed.ncbi.nlm.nih.gov/32130618>
LncRNA SNHG7 promotes cardiac remodeling by upregulating ROCK1 via sponging miR-34-5p	2020	endomyocardial_fibrosis	<https://pubmed.ncbi.nlm.nih.gov/32507765>
Fendrr involves in the pathogenesis of cardiac fibrosis via regulating miR-106b/SMAD5 axis	2020	heart_failure	<https://pubmed.ncbi.nlm.nih.gov/31982134>
LncRNA SOX2OT/Smad3 feedback loop promotes myocardial fibrosis in heart failure	2020	heart_failure	<https://pubmed.ncbi.nlm.nih.gov/32959533>
LncRNA H19 ameliorates myocardial infarction-induced myocardial injury and maladaptive cardiac remodelling by regulating KDM3A	2020	myocardial_infarction	<https://pubmed.ncbi.nlm.nih.gov/31755219>
Mechanism underlying increased cardiac extracellular matrix deposition in perinatal nicotine-exposed offspring	2020	endomyocardial_fibrosis	<https://pubmed.ncbi.nlm.nih.gov/32795172>
Mechanism underlying increased cardiac extracellular matrix deposition in perinatal nicotine-exposed offspring	2020	ventricular_dysfunction	<https://pubmed.ncbi.nlm.nih.gov/32795172>
Long noncoding RNA-GAS5 retards renal fibrosis through repressing miR-21 activity	2020	kidney_failure_chronic	<https://pubmed.ncbi.nlm.nih.gov/32791063>
LncRNA GAS5 exacerbates renal tubular epithelial fibrosis by acting as a competing endogenous RNA of miR-96-5p	2020	diabetic_nephropathy	<https://pubmed.ncbi.nlm.nih.gov/31810148>
Long noncoding RNA NEAT1 is involved in the protective effect of Klotho on renal tubular epithelial cells in diabetic kidney disease through the ERK1/2 signaling pathway	2020	diabetic_nephropathy	<https://pubmed.ncbi.nlm.nih.gov/32054986>
Silencing of the lncRNA TUG1 attenuates the epithelial-mesenchymal transition of renal tubular epithelial cells by sponging miR-141-3p via regulating beta-catenin	2020	nephritis_interstitia	<https://pubmed.ncbi.nlm.nih.gov/33135476>
LncRNA KCNQ1OT1 contributes to cardiomyocyte apoptosis by targeting FUS in heart failure	2020	heart_failure	<https://pubmed.ncbi.nlm.nih.gov/32497620>
Long noncoding RNA NEAT1 sponges miR-130 to modulate renal fibrosis by regulation of collagen type I	2020	kidney_diseases	<https://pubmed.ncbi.nlm.nih.gov/32475133>
HIF-1alfa-upregulated lncRNA-H19 regulates lipid droplet metabolism through the AMPKalfa pathway in hepatic stellate cells	2020	liver_cirrhosis	<https://pubmed.ncbi.nlm.nih.gov/32445757>
LncRNA H1cal2/miR-203a-3p sponge participates in epithelial-mesenchymal transition by targeting p66Shc in liver fibrosis	2020	liver_cirrhosis	<https://pubmed.ncbi.nlm.nih.gov/32659284>
Long non-coding RNA GAS5 regulates myocardial ischemia-reperfusion injury through the PI3K/AKT apoptosis pathway by sponging miR-532-5p	2020	cardiovascular_diseases	<https://pubmed.ncbi.nlm.nih.gov/31985016>

Figure 14: Query 5 output.

Results	Execution Time
16	1sec, 273ms

Figure 15: Query 5 results.

4.2. R queries

The R script with the queries to sparql is in *query_r_sparql.R*

5. Publication of the RDF database

The RDF database and metadata were with published on Trifid, with the URL:

http://dayhoff.inf.um.es:8176/dataset/Dataset_lncrna_mus

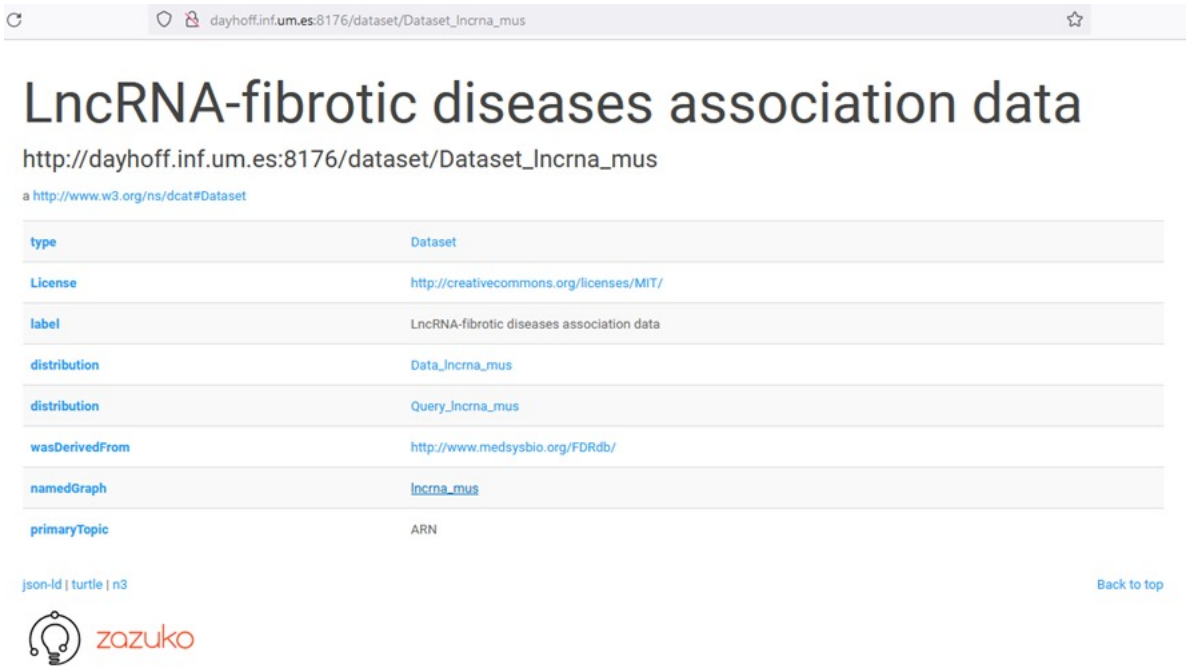

Figure 16: Screenshot from the uploaded dataset on the Trifid interface.