Biocuration of microRNAs in idiopathic pulmonary fibrosis

Yalbi I. Balderas-Martínez[1], Fabio Rinaldi[2], Gabriela Contreras[3], Hilda Solano[3], Mishael Sánchez-Pérez[3], Socorro Gama-Castro[3], Julio Collado-Vides[3], Moisés Selman[4], and Annie Pardo[1]

[1] Facultad de Ciencias, UNAM, México,D.F., MX
yalbibalderas@ciencias.unam.mx
[2] Institute of Computational Linguistics, University of Zurich,Zurich.,SE
[3] Computational Genomics Program,Center for Genomics Science,UNAM, Cuernavaca,Morelos.,MX
[4] Instituto Nacional de Enfermedades Respiratorias, México, D.F., MX

Abstract. MicroRNAs (miRNAs) are small and non-coding RNA molecules that inhibit gene expression post-transcriptionally. They play important roles in several biological processes, and in recent years, there have been an interest in studying how they are related to the pathogenesis of respiratory diseases. Although there are already some databases that do it, curating miRNAs is a big challenge due to the amount of information that is being generated everyday. Respiratory diseases are poorly documented in databases, in spite of the fact that they are of increasing concern in terms of morbidity, mortality, and economic impact.
As part of our participation in the IAT track of BioCreative V, we adapted the OntoGene text mining pipeline and the ODIN curation system to the task of curating miRNAs in relation to one particular respiratory disease, idiopathic pulmonary fibrosis, the most common of the interstitial pneumonias. It is a chronic disease of unknown cause, irreversible, progressive and lethal. We curated almost 300 miRNAs related to this disease using a semi-automatic approach with the system OntoGene/ODIN. Our text mining protocol can be applied to obtain the miRNAs of all the respiratory diseases and with the possibility to expand to other diseases.
Separately the OntoGene/ODIN curation system was also made available to independent curators contacted by the IAT track organizers, who all commented positively on the capabilities and usability of the sytem.

Key words: miRNA, IPF, Text mining, Biocuration, Lung Disease, Ontogene, ODIN

1 Introduction

MicroRNAs (miRNAs) are small and non-coding RNA molecules that inhibits gene expression post-transcriptionally [1]. They play crucial roles in important

biological processes, in recent years, there have been an interest in studying how they are related to the pathogenesis of human diseases. Although they have not been discovered so recently [2], the term microRNA was introduced only in 2001 [3], so it is not trivial to find papers associated with miRNAs before this date. After that, the number of experimental research has been increased exponentially, see Fig 1.

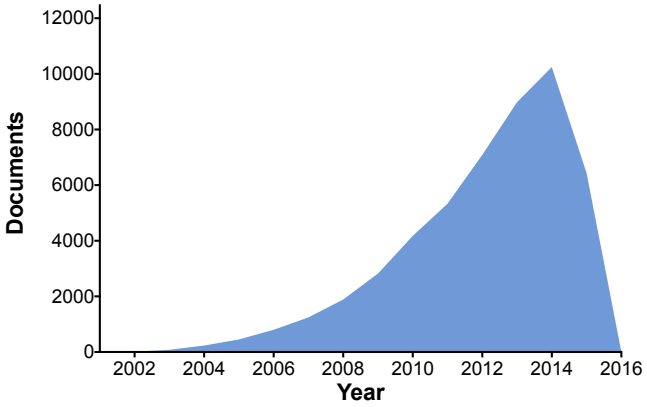

Fig. 1. Number of publications related with the term microRNA in SCOPUS database (Source: http://www.scopus.com/).

At this moment, in human, there are about 38,113 experimentally validated interactions (between miRNA & its target gene) for 587 miRNAs in a total of 2,143 papers in mirTarBase v4.0, an experimentally validated microRNA-target interactions database [4]. However, some studies have suggested that miRNAs could regulate about 60% of the human genome [5], meaning that the number of interactions will increase, and this could represent a major problem to add the information in databases. So, although there is an interest in curation of the miRNAs associated to diseases, this represents a big challenge due to the amount of information that is being generated everyday. For instance, in miRTarBase, there are only ten respiratory diseases represented with only few miRNAs associated, in spite of the fact that respiratory diseases are of increasing concern in terms of morbidity, mortality, and economic impact [6].

For the previous reason, we are interested in having a semi-automatic approach to curate all the interactions associated to respiratory diseases, but considering that this is a pilot project, we would like to start with only one disease: idiopathic pulmonary fibrosis (IPF). IPF is the most common of the interstitial pneumonias, is chronic, irreversible, progressive and lethal of unknown cause [7]. It is known that many miRNAs are differentially expressed in people with IPF *vs.* healthy as it has been noted in the most recent review published [8].

The BioCreative series of competitive evaluations of biomedical text mining technologies have inspired the development of novel text mining tools that have real world application. The OntoGene text mining system, originally developed for the extraction of entities and their interactions, proved to be quite successful in previous BioCreative challenges (e.g. best results in finding protein-protein interactions in BioCreative 2009 [9]). Since 2010, OntoGene is available in a client-server architecture where the remote user can access the results of the annotation pipeline through the browser-based system interface (ODIN: OntoGene Document Inspector), which then can access the text mining services residing on the remote server. OntoGene/ODIN has already been used in the curation of different biological elements with reliable results [10]. We decided to participate in the User Interactive Task of Biocreative V challenge to evaluate how this tool can improve the biocuration process for miRNAs, comparing the manual *vs* a semi-automated approach. In the next section we describe the approach used for the miRNA curation, and briefly mention the additional curation tasks that were offered to IAT curators.

2 Methods

2.1 Obtaining the dataset

We found 63 papers in PubMed associated with the terms "microRNAs", "idiopathic pulmonary fibrosis" and filtering only for the human ones. We excluded one paper because it was referred to a different disease. So, the final corpus contained 62 papers. Only free full length articles were converted to text and then added to ODIN (http://www.ontogene.org).

2.2 Creating the annotation dictionary

We are interested in obtaining the following information:

- MicroRNA name
- Target genes affected
- Transcription factors associated
- Organism (human)
- Disease associated (idiopathic pulmonary fibrosis)
- Level of the microRNA in some condition (up, down, overexpressed, deleted, induced, repressed, etc)
- Some characteristics of the samples:
  - Type of sample (lung tissue, alveolar macrophages, fibroblasts, etc)
  - Gender (Female, Male)
  - Age
  - Condition (Healthy vs IPF)
  - Smoking status (Current, former, past)
  - Race (Caucasian, afroamerican, etc)

To find this information we have created a list of terms to be used by the OntoGene/ODIN system as an annotation dictionary. First, we have added the names of all the microRNAs that are known in mirTarBase [4], we included all the names for the different species, in case that they could be found also in humans. Second, we have added the names of all the transcription factors known in the database for eukaryotes Jaspar [11]; and then, all the names of the human genes using HUGO nomenclature [12] . Finally, we have selected some terms related to the characteristics of the samples and gene regulation with the help of an expert committee. All the terms have been classified into some categories, or term types (microRNA, Disease, Organism, Sample, Gene, Transcription Factor, etc), to facilitate the use of ODIN filters.

2.3 Additional Tasks

Curators contacted by the IAT track organizers were given the option to choose among three versions of OntoGene/ODIN:

- CTD version
  Bioconcepts: gene / disease / chemicals
  Standards: CTD
  Information Extraction: gene / disease / chemicals, and their interactions
  Text: abstracts
  `http://kitt.ifi.uzh.ch/kitt/ODIN/bc2015/`
- RegulonDB version
  Bioconcepts: genes, transcription factors, methods, conditions, effects, and other RegulonDB concepts
  Standards: RegulonDB
  Information Extraction: bionconcepts mentioned before, no relations are automatically computed, but curators can use filters to detect co-occurences of specific entity types, and thus curate relationships
  Text: full text
  `http://kukulcan.ccg.unam.mx/~ontogene/ODIN/bc2015-ccg/`
- miRNA version
  Bioconcepts: miRNA, gene, transcription factor, and sample characteristics in respiratory diseases.
  Standards: mirTarBase, Jaspar
  Information Extraction: bionconcepts mentioned before, no relations are automatically computed, but curators can use filters to detect co-occurences of specific entity types, and thus curate relationships, in particular positive and negative regulation of trascription factors, and target genes of miRNAss
  Text: full text
  `http://kukulcan.ccg.unam.mx/~ontogene/ODIN/bc2015-miRNA/`

Detailed user documentation was provided in the form of a user manual and screencasts detailing the most important functionalities of the system. The screencasts describe in particular the CTD version of the system, but they are

Fig. 2. Example of phrases highlighted in an article using ODIN

usable to a large extent also for the other two versions. Videos and documentation are accessible from a link within the ODIN system. The CTD version is trained on CTD datasets but can be applied on any PubMed abstract. The RegulonDB version uses a set of articles provided by the Regulon Database. The miRNA version is described in the rest of this paper.

3 Results and Discussion

3.1 miRNA task

ODIN provides an interface in which the biocurator can read only the relevant phrases of the paper using some filters. As "relevant", we mean to display a term highlighted if it appears in the dictionary that we have previously created. We have applied filters, to show only those phrases that contain a term type, for example, in the PMID 24242359 [13], we obtained 23 phrases using the filter "microRNA", that contains a list of miRNAs, although nine of these phrases were located in the section of references. However, not all the phrases that contain miRNAs are important in IPF. If we add another filter, like disease, we obtain what we were looking for, see Figure 2.

In this example, we observe there are many terms highlighted that corresponds to: MicroRNAs name: miR-let7d, miR-200, miR-31, miR-17, let-7d; Disease associated: IPF; MicroRNA level: downregulated, downregulation, overexpression; Target gene: HMGA2; Sample type: alveolar epithelial cells.

It is important to mention that the characteristics of the samples are difficult to obtain in most of the papers because they are frequently shown in figures, tables or supplementary material, and at this moment we have not implemented a protocol to work with these sections, so they were curated manually.

In addition we depend on the miRNAs already annotated in databases, but it is also important to look for those that are not contained in any databases.

Therefore we could extend our results by applying filters using the term "microRNA" and the name of a gene, which would allow us to find new miRNAs and their targets.

Besides, to truly understand and annotate the interactions, it is necessary to have a curator to read the phrases. Based on our experience in manual curation we know that this semi-automatic process allow us to save time, since we have to read only the phrases that contain the information we need. A curator read about one paper per day, this means that the total curation is completed in 62 days. With ODIN this work was reduced to only five days.

Finally, using this semi-automatic approach we could curate almost 300 hundred of microRNAs associated to the idiopathic pulmonary fibrosis, so we expect a promising result when we apply this protocol to other respiratory diseases.

3.2 Other tasks

As mentioned above, curators invited by the IAT organizers were free to chose one of the three ODIN versions described in section 2.3. The IAT organizers asked them later to fill in a questionnaire with questions aimed at verifying their perception of the usability of the system based on their experience. Based on their answers, OntoGene/ODIN scored 91.66% on the System Usability Scale (SUS), with a value of Usability of 90.62% and a value of Learnability of 95.83%. ODIN also scored very highly on the metrics of task completion and design.

Acknowledgements YIB-M., acknowledges the program of postdoctoral fellowship DGAPA-UNAM and the Faculty of Sciences Universidad Nacional Autónoma de México. The development of OntoGene/ODIN has been supported by the Swiss National Science Foundation (grant 105315130558/1, PI: Fabio Rinaldi) and by the Data Science Group at Hoffmann-La Roche, Basel, Switzerland. MS-P acknowledges DGAPA-UNAM-CCG for the postdoctoral fellowship.

References

1. Rupani, H., Sanchez-Elsner, T., Howarth, P.: MicroRNAs and respiratory diseases. Eur Respir J **41**(3) (Mar 2013) 695–705
2. Lee, R.C., Feinbaum, R.L., Ambros, V.: The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75**(5) (1993) 843 – 854
3. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T.: Identification of novel genes coding for small expressed RNAs. Science **294**(5543) (2001) 853–858
4. Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., Jian, T.Y., Lin, F.M., Chang, T.H., Weng, S.L., Liao, K.W., Liao, I.E., Liu, C.C., Huang, H.D.: MiRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res **42**(Database issue) (Jan 2014) D78–85
5. Friedman, R.C., Farh, K.K.H., Burge, C.B., Bartel, D.P.: Most mammalian mRNAs are conserved targets of microRNAs. Genome Res **19**(1) (Jan 2009) 92–105

6. Speizer, F.E., Horton, S., Batt, J., Slutsky, A.S.: Respiratory diseases of adults. 2 edn. World Bank (2006)
7. King, T.E.J., Pardo, A., Selman, M.: Idiopathic pulmonary fibrosis. Lancet **378**(9807) (Dec 2011) 1949–1961
8. Pandit, K.V., Milosevic, J.: MicroRNA regulatory networks in idiopathic pulmonary fibrosis. Biochem Cell Biol **93**(2) (Apr 2015) 129–37
9. Rinaldi, F., Schneider, G., Kaljurand, K., Clematide, S., Vachon, T., Romacker, M.: OntoGene in BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics **7**(3) (2010) 472–480
10. Gama-Castro, S., Rinaldi, F., López-Fuentes, A., Balderas-Martínez, Y.I., Clematide, S., Ellendorff, T.R., Santos-Zavaleta, A., Marques-Madeira, H., Collado-Vides, J.: Assisted curation of regulatory interactions and growth conditions of oxyr in e. coli k-12. Database **2014** (2014)
11. Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W.W.: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res **42**(Database issue) (Jan 2014) D142–7
12. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A.: Genenames.org: the HGNC resources in 2015. Nucleic Acids Res **43**(Database issue) (Jan 2015) D1079–85
13. Liu, Y., Li, H., Xiao, T., Lu, Q.: Epigenetics in immune-mediated pulmonary diseases. Clin Rev Allergy Immunol **45**(3) (Dec 2013) 314–30