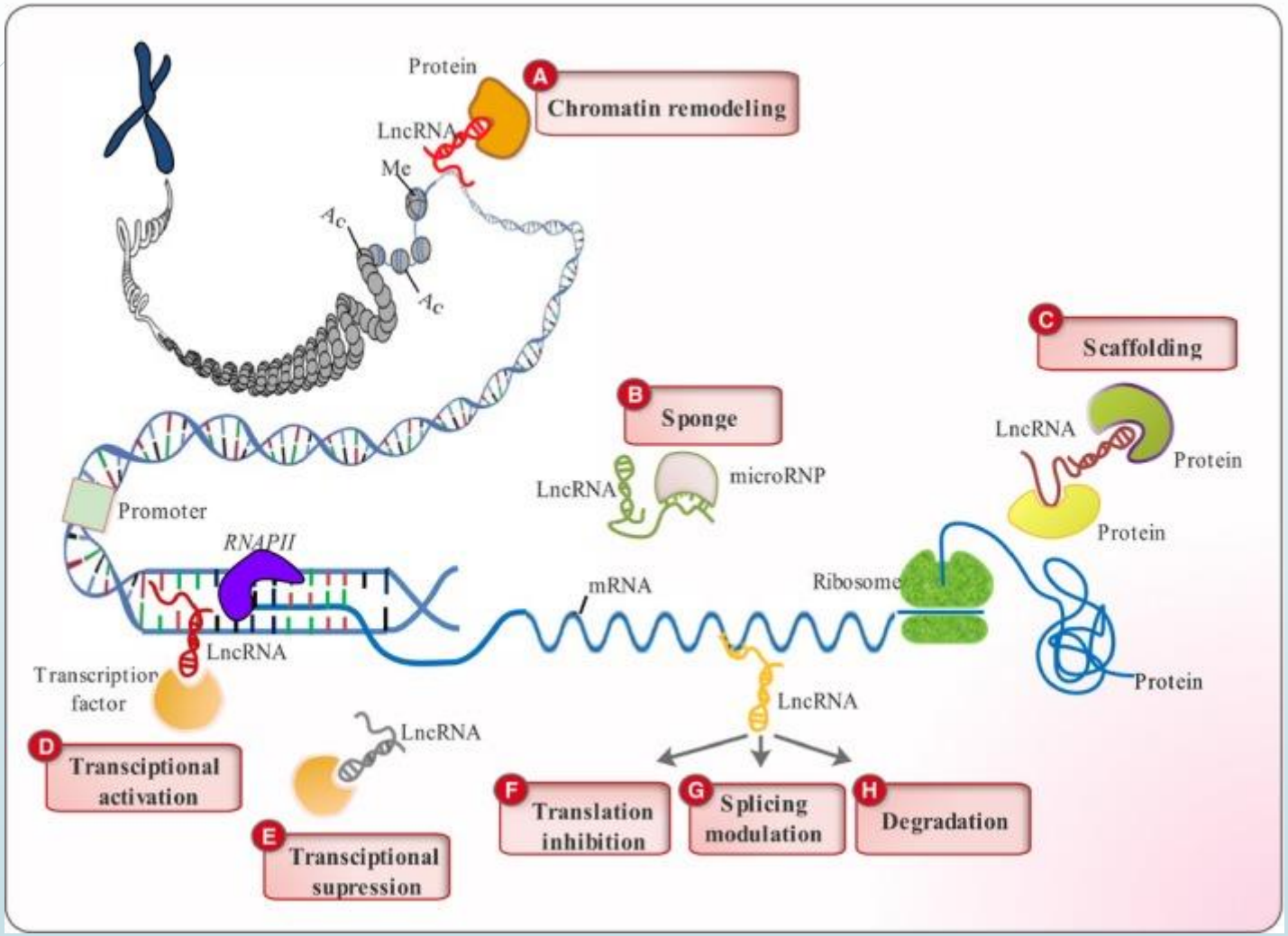
Neighborhood based approaches for the prediction of lncRNA-Disease association from tripartite graphs

Ricercatori coinvolti: **Mariella Bonomo**, Armando La Placa, Simona E. Rombo

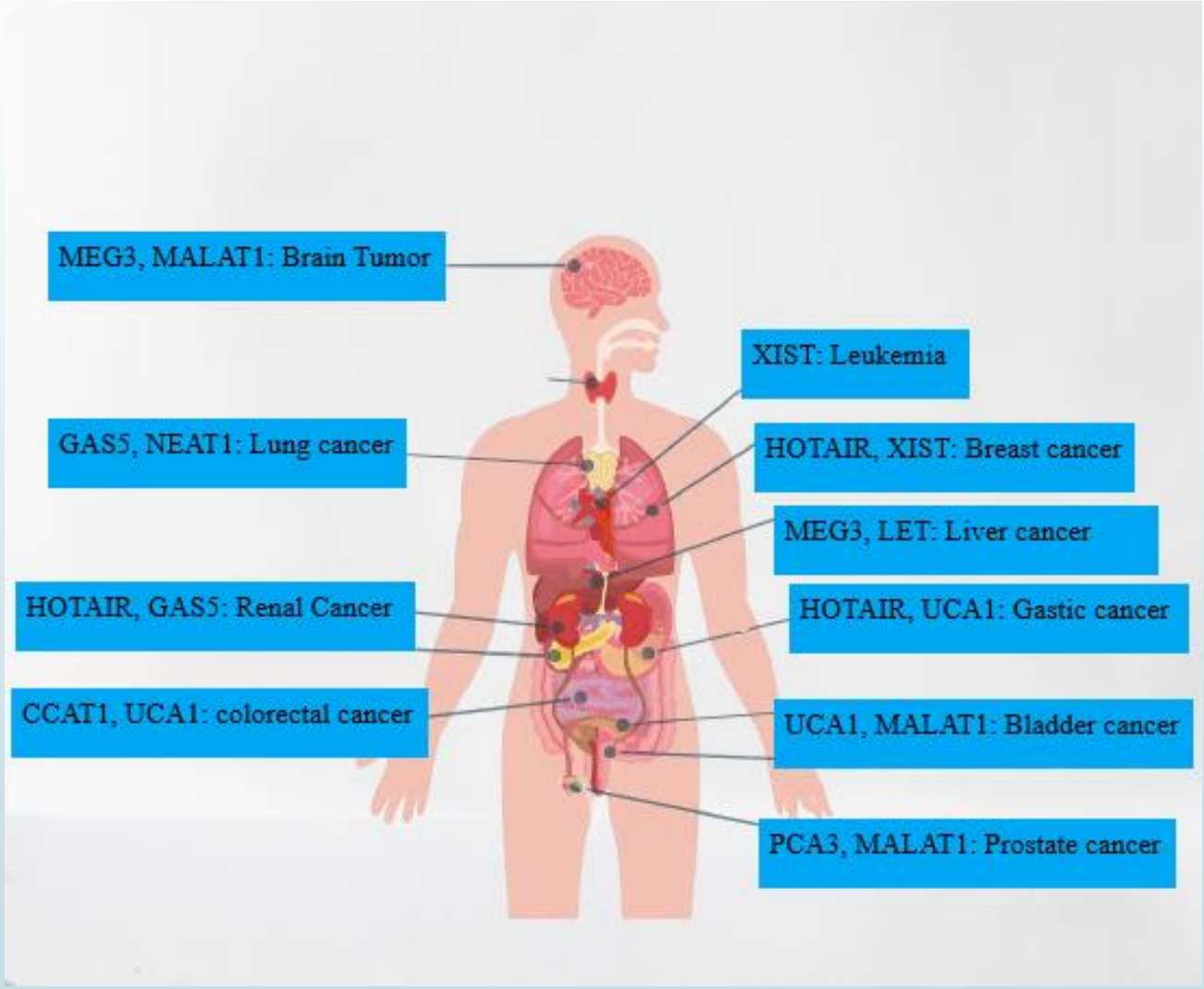
Università degli Studi di Palermo

Functions of Long non-coding RNAs



Taken from "State of The Art Technologies Used to Explore Long Non-coding RNAs in Cancer Disease", S. Salehi et al., Stat. J. Cell. Mol. Med., 20(10):1–21, 2017

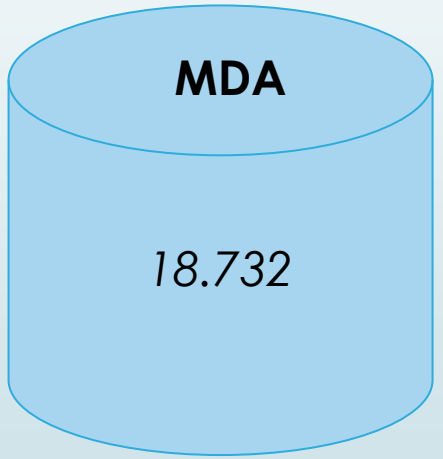
LncRNAs have a role in human diseases



Problem

- Only a few experimentally verified lncRna – disease associations are known

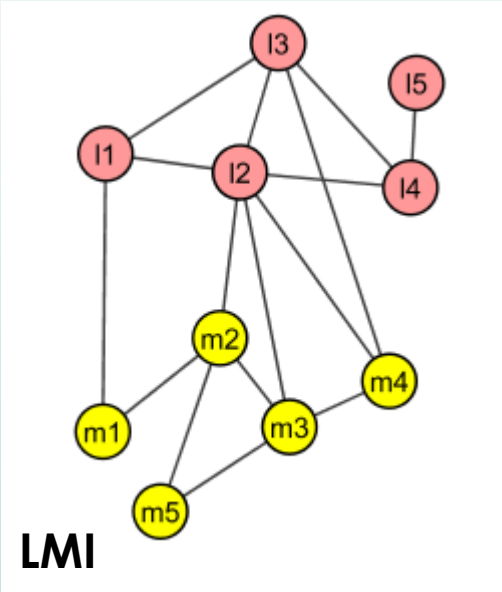
LncRNA – disease associations stored in cuiLab db



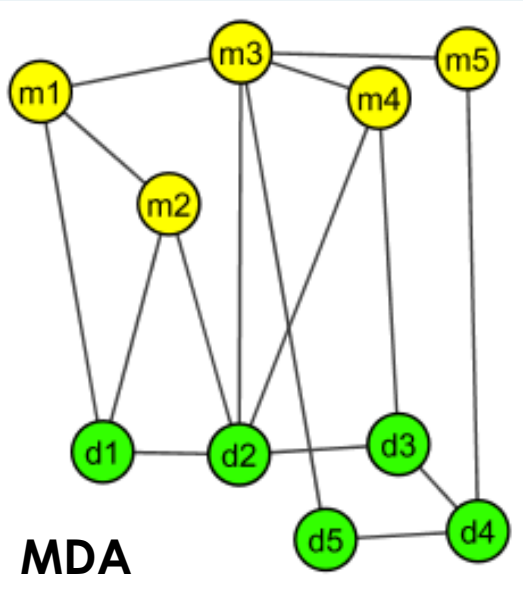
miRNA – disease associations stored in *HMDD db*

Observation

- Large amounts of lncRNA-miRNA interactions and miRNA-disease associations have been collected in public databases



10.112 lncRNA-miRNA in starBase db



18.732 miRNA-disease associations In HMDD db

Computational approaches for lncRNA-disease associations prediction

Do not use known LDA

X.Chen., 2015, Scientific Reports

S. Alaimo et al. , 2014, Frontiers in Bioengineering and Biotechnology

Do use known LDA

C. Domeniconi et al. , 2018, Bioinformatics

M.N.Wang et al. 2021, Neurocomputing

……..

……..

Computational approaches for lncRNA-disease associations prediction

Do not use known LDA

X.Chen., 2015, Scientific Reports

S. Alaimo et al. , 2014, Frontiers in Bioengineering and Biotechnology

Do use known LDA

C. Domeniconi et al. , 2018, Bioinformatics

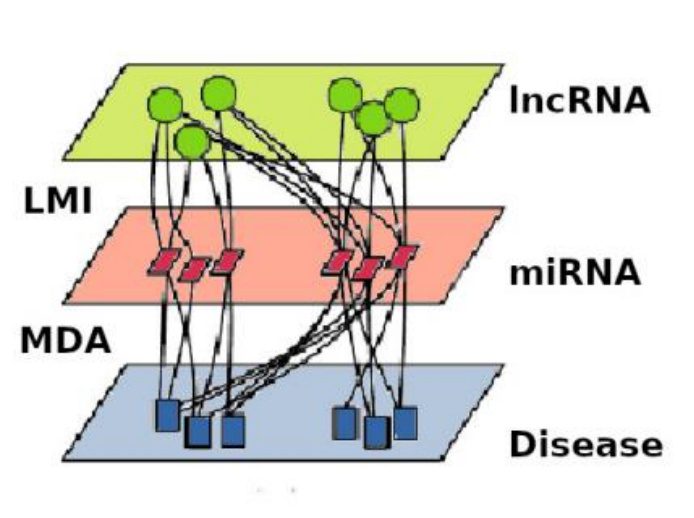
M.N.Wang et al. 2021, Neurocomputing

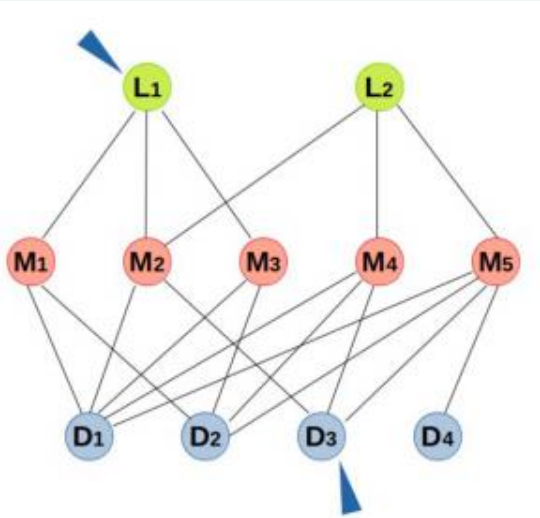
Proposed Approach

- Predict lncRNA-disease associations (LDAs) based on known lncRNA-miRNA interactions (LMIs) and miRNA-disease associations (MDAs)
- Consider miRNAs as intermediate molecules

Tripartite Graph Representation

Model LDAs, LMIs and MDAs as a tripartite graph





lncRNA 1.114
miRNA 1.065
diseases 885

The goal is to return a set of predicted LDAs

Neighborhood based approach

- For each pair lncRNA-disease (l,d) return in output an LDA with score:

$$S(l_i, d_j) = \alpha \cdot \frac{|M_{l_i} \bigcap M_{d_j}|}{|M_{l_i} \bigcup M_{d_j}|} + (1 - \alpha) \cdot \frac{|\bigcup_x (M_{l_x} \bigcap M_{d_j})|}{|\bigcup_x (M_{l_x}, M_{d_j})|}$$

- measures how much "connected“ L and D are in the tripartite graph

M_{li} contains the number of miRNAs associated to l_i

M_{dj} contains the number of miRNAs associated to d_j

Method

- Let T_{LMD} a tripartite graph
- We defined a prediction score LDA, based on neighborhood analysis with Apache Spark
- Statistical test based on recent experimental literatures : False Discovery Rate (FDR)

Datasets and gold standard

mirna-disease association (**MDA 18.732**) from HMDD db with:

- **1206 miRNAs and 894 diseases**

lncRNA-mirna interaction (**LMI 10.112**) from starBase db, with:

- 132 miRNAs and 1114 lncRNAs

lncRNA-disease association (**LDA 1564**) from cuilab db with:

- 914 lncRNA and 329 disease

A gold standard dataset from lncRNA disease db (X. Chen 2015) with:

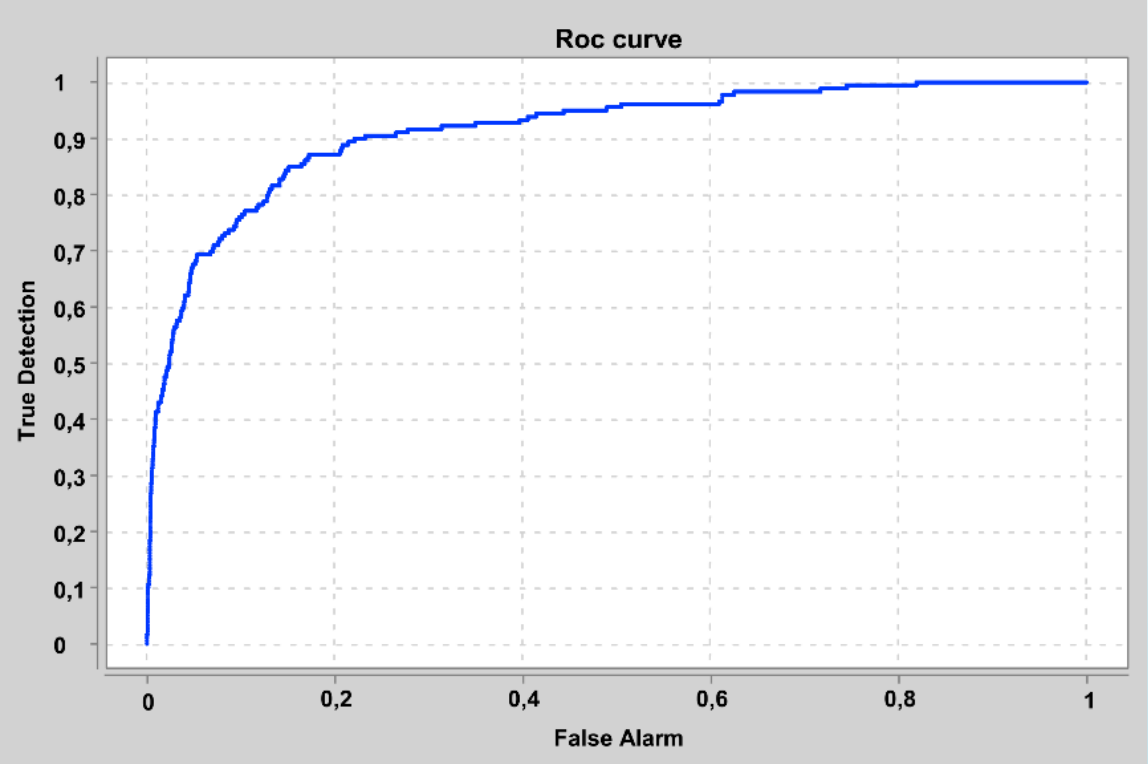
- 183 lncRNA disease

LOOCV:

- The predicted LDAs are ranked according to their *corrected* score
- Each verified LDA is left out in turn as test sample
- When the rank of this test sample exceeds a given threshold, the model provides a successful prediction
- At the varying of the threshold, compute true positive rate (TPR, sensitivity) and false positive rate (FPR, specificity)
  - Sensitivity: % of the test samples whose ranking is higher than the given threshold
  - Specificity: % of samples that are below the threshold
- We evaluated test by curve analysis ROC

ROC curve

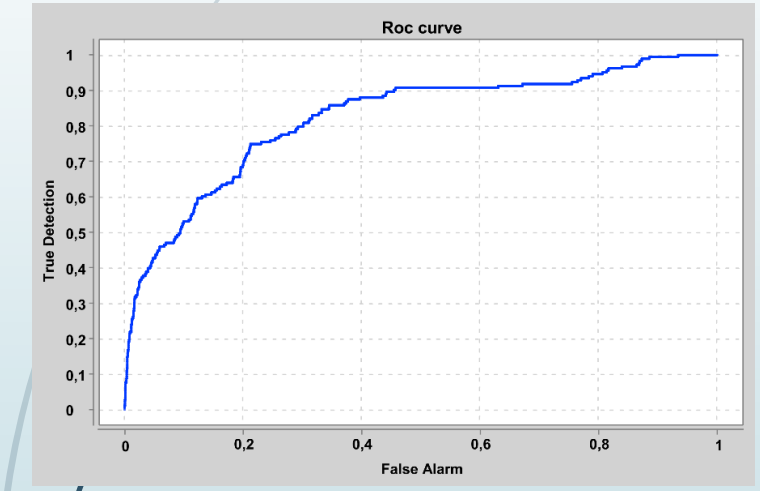
- Receiver-Operating Characteristics (ROC) curve is drawn by plotting TPR versus FPR at different thresholds



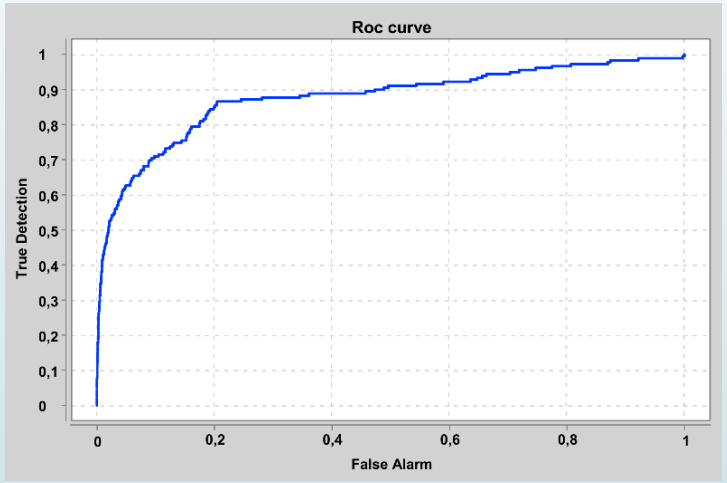
AUC = 0.91

Results with same datasets

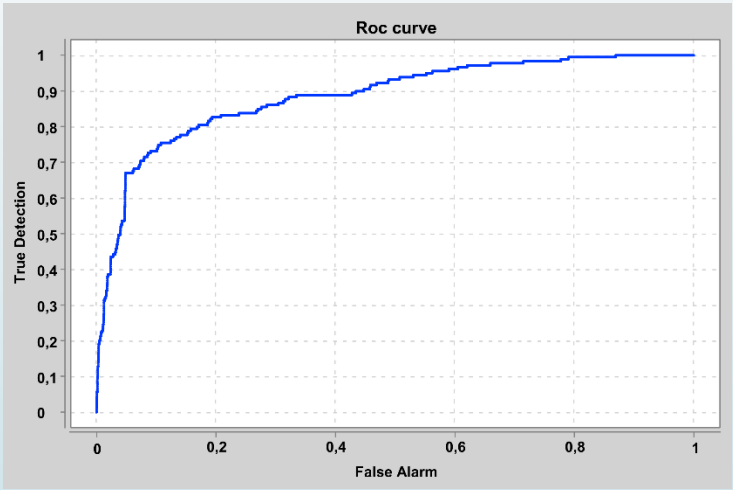
HMDD v3 dataset



Bonomo et al.
AUC = 0.91



Chen et al.
AUC = 0.87



Alaimo et al.
AUC = 0.88

Work in Progress

Do not use known LDA

X.Chen., 2015, Scientific Reports

S. Alaimo et al. , 2014, Frontiers in Bioengineering and Biotechnology

Do use known LDA

C. Domeniconi et al. , 2018, Bioinformatics

M.N.Wang et al. 2021, Neurocomputing

Work in Progress

Do not use known LDA

X.Chen., 2015, Scientific Reports

S. Alaimo et al. , 2014, Frontiers in Bioengineering and Biotechnology

Do use known LDA

C. Domeniconi et al. , 2018, Bioinformatics

M.N.Wang et al. 2021, Neurocomputing

…….

Related Publications

- Bonomo M.; La Placa A.; Rombo S.E. Prediction of lncRNA-Disease Associations from Tripartite Graphs. DOI:10.1007/978-3-030-71055-2_16. pp.205-210. In VLDB Workshops, Poly 2020 and DMAH 2020, Virtual Event, August 31 and September 4, 2020. In LECTURE NOTES IN COMPUTER SCIENCE - ISSN:1611-3349 vol. 12633

- Bonomo M.; La Placa A.; Rombo S.E. Prediction of Disease–lncRNA Associations via Machine Learning and Big Data Approaches. In Knowledge Modelling and Big Data Analytics in Healthcare, CRC Press, 2021.

Thank you!