UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

dib DIPARTIMENTO DI INFORMATICA

KDDE

Multi-type Clustering for the Identification of lncRNA-disease Relationships

Gianvito Pio, Francesco Serafino, Emanuele Barracchia, Domenica D’Elia, Donato Malerba and Michelangelo Ceci

Dipartimento di Informatica, Università degli Studi di Bari “Aldo Moro”, Bari, Italy

CNR- Institute for Biomedical Technologies, Bari, Italy

National Research Council of Italy

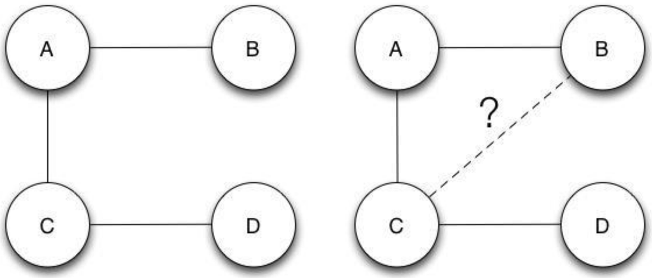
L A C A M

Knowledge Discovery & Data Engineering

MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Introduction



ncRNA/DNA or mRNA interactions are currently **hot topics** in biology

- The first influence the latter by modulating their expression capacity
- non-coding RNAs can be:
  - **short** non-coding RNA (e.g., miRNAs)
  - **long** non-coding RNA (lncRNAs, > 200nt long)

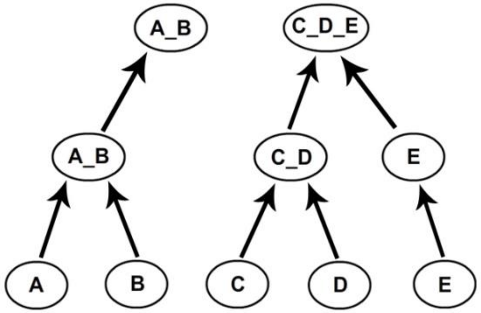
lncRNAs are also responsible of **diseases**

- **poor** amount of experimentally **evaluated interactions**
- **only positive examples** of interactions are available

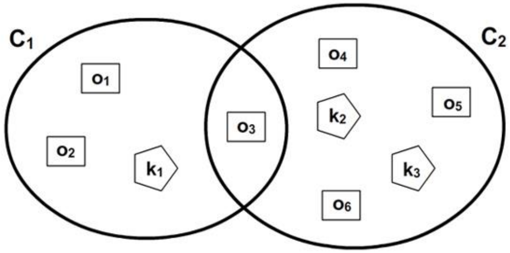
Multi-type Clustering

Idea: identification of heterogenous clusters of lncRNAs and diseases from network data (lncRNAs, miRNAs, target genes, diseases)

- Hierarchical Clustering



- Multi-type and overlapping



- Facilitates the understanding of results by human experts
- Allows choosing globally-based or locally-based predictions
- Catches relationships among objects of multiple types, which can be involved in different sub-networks

Prediction of relationships

1. Extraction of **experimentally validated** interactions from the network
2. Extraction of a **hierarchy of multi-type clusters** by applying our multi-type clustering algorithm
3. Generation of all the possible **new interactions**
4. For each hierarchical level:
   - Determine a **score** for each unknown interaction as the **combination** of the cohesiveness values associated with the clusters it belongs to
   - Return **new interactions** with their associated **score**

Cohesiveness: average strength of the relationships in the cluster

Computation of the prediction score

- Let $s(t_i\ )$ be the score associated to the pair t_i
- Let G_{t_i} be the set of clusters t_i belongs to
  - If $|G_{t_i}| = 1$ and $G' \in G_{t_i}$, $s(t_i\ ) = cohes(G')$
  - Otherwise, we consider some combination strategies:
    1. **Max:** $s(t_i) = \max_{G_j{}' \in G_{t_i}} cohes(G_j')$
    2. **Min:** $s(t_i) = \min_{G_j{}' \in G_{t_i}} cohes(G_j')$
    3. **Avg:** $s(t_i) = \frac{1}{|G_{t_i}|} \sum_{G_j{}' \in G_{t_i}} cohes(G_j')$
    4. **Custom:** $s(t_i) = \boldsymbol{f}(G_{t_i})$

The proposed custom combination strategy aims at rewarding those interactions falling in multiple (highly-cohesive) clusters.

Experiments - Evaluation

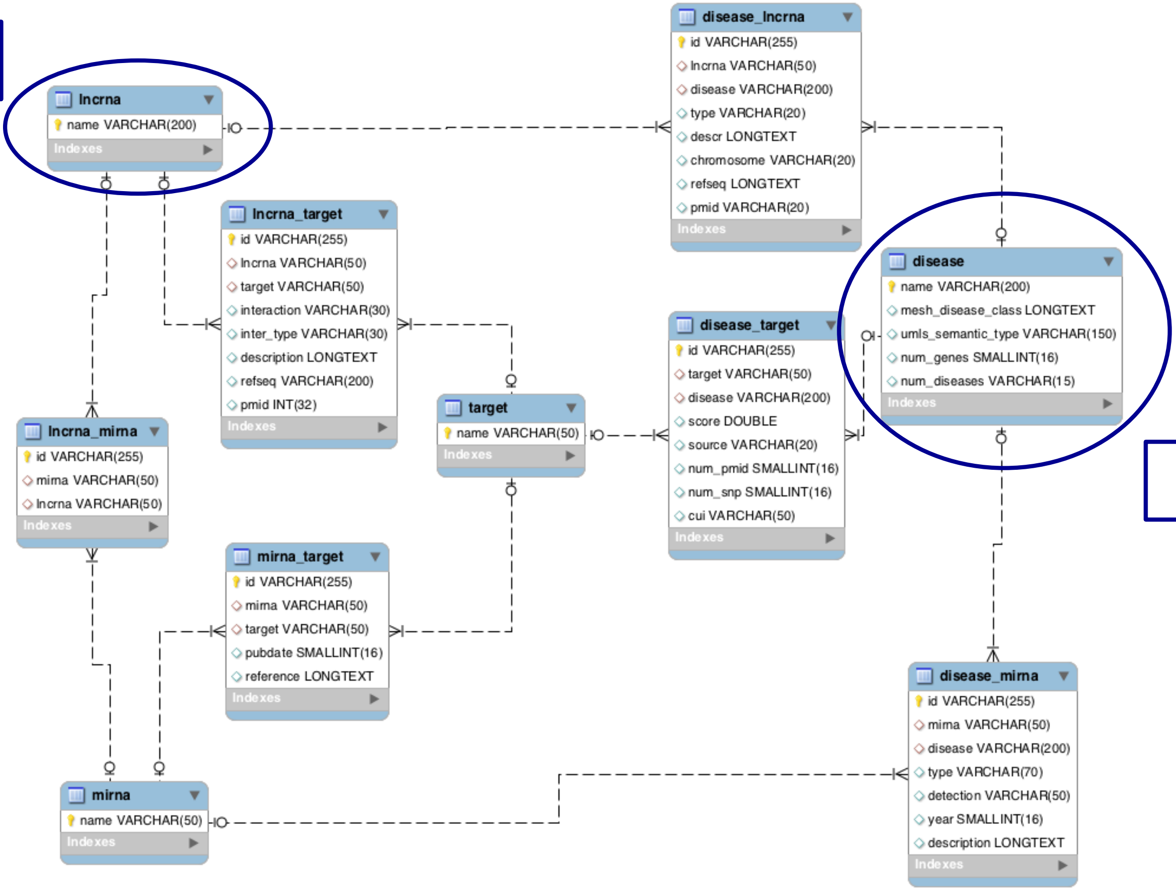
- Systems
  - LP-MTRCLUS (Our System)
  - HOCCLUS2
- Measure
  - True Positive Rate $TPR = \frac{TP}{TP+FN}$
- Evaluation performed
  - By moving a threshold on the prediction scores
  - On the first three hierarchical levels

Experiments - Dataset

- We built an heterogeneous network starting from:
  - "Base Pairing between miRNAs and Their Non-mRNA Targets" [Helwak et al.] (miRNA:lncRNA)
  - "lncRNA interaction data" [Chen et al. 2013] (lncRNA:target_genes)
  - "lncRNA-disease association data" [Chen et al. 2013] (lncRNA:diseases)
  - DisGeNET (diseases:target_genes)
  - Mir2Disease [Jiang et al. 2009] (miRNA:target_genes, miRNA:diseases)

Experiments - Dataset

lncrna



lncrna
- name VARCHAR(200)
- Indexes

disease_lncrna
- id VARCHAR(255)
- lncrna VARCHAR(50)
- disease VARCHAR(200)
- type VARCHAR(20)
- descr LONGTEXT
- chromosome VARCHAR(20)
- refseq LONGTEXT
- pmid VARCHAR(20)
- Indexes

lncrna_target
- id VARCHAR(255)
- lncrna VARCHAR(50)
- target VARCHAR(50)
- interaction VARCHAR(30)
- inter_type VARCHAR(30)
- description LONGTEXT
- refseq VARCHAR(200)
- pmid INT(32)
- Indexes

disease
- name VARCHAR(200)
- mesh_disease_class LONGTEXT
- umls_semantic_type VARCHAR(150)
- num_genes SMALLINT(16)
- num_diseases VARCHAR(15)
- Indexes

disease_target
- id VARCHAR(255)
- target VARCHAR(50)
- disease VARCHAR(200)
- score DOUBLE
- source VARCHAR(20)
- num_pmid SMALLINT(16)
- num_snp SMALLINT(16)
- cui VARCHAR(50)
- Indexes

target
- name VARCHAR(50)
- Indexes

lncrna_mirna
- id VARCHAR(255)
- mirna VARCHAR(50)
- lncrna VARCHAR(50)
- Indexes

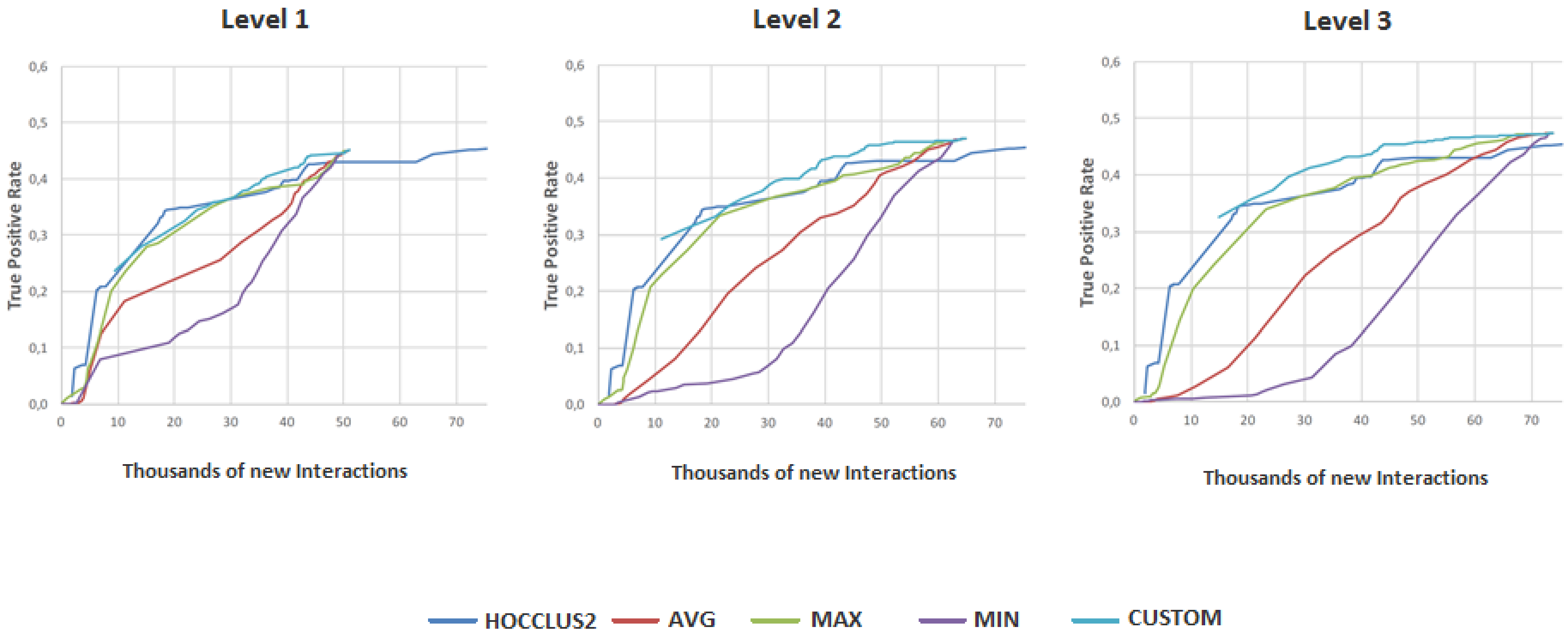
mirna_target
- id VARCHAR(255)
- mirna VARCHAR(50)
- target VARCHAR(50)
- pubdate SMALLINT(16)
- reference LONGTEXT
- Indexes

mirna
- name VARCHAR(50)
- Indexes

disease_mirna
- id VARCHAR(255)
- mirna VARCHAR(50)
- disease VARCHAR(200)
- type VARCHAR(70)
- detection VARCHAR(50)
- year SMALLINT(16)
- description LONGTEXT
- Indexes

disease

Results


Level 1
Level 2
Level 3
True Positive Rate
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0
10
20
30
40
50
60
70
Thousands of new Interactions
HOCCLUS2
AVG
MAX
MIN
CUSTOM

Conclusions and Future Works

- Multi-type clustering can be **fruitfully exploited for Link Prediction** purposes
- The proposed strategy for combining the scores allows us to **perform accurate predictions** that **outperform competitors and baseline approaches**

- As future work we intend to evaluate our system:
  - On interactions among objects of **more than 2 different types**
  - By considering the possible **type of relationships** among objects

Questions: tomorrow 10:20-11:20 – Poster P7

Thanks for your attention

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944) and CHARME

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA