KU LEUVEN kulak

LncRNA disease association prediction using multi-label approaches

Introduction

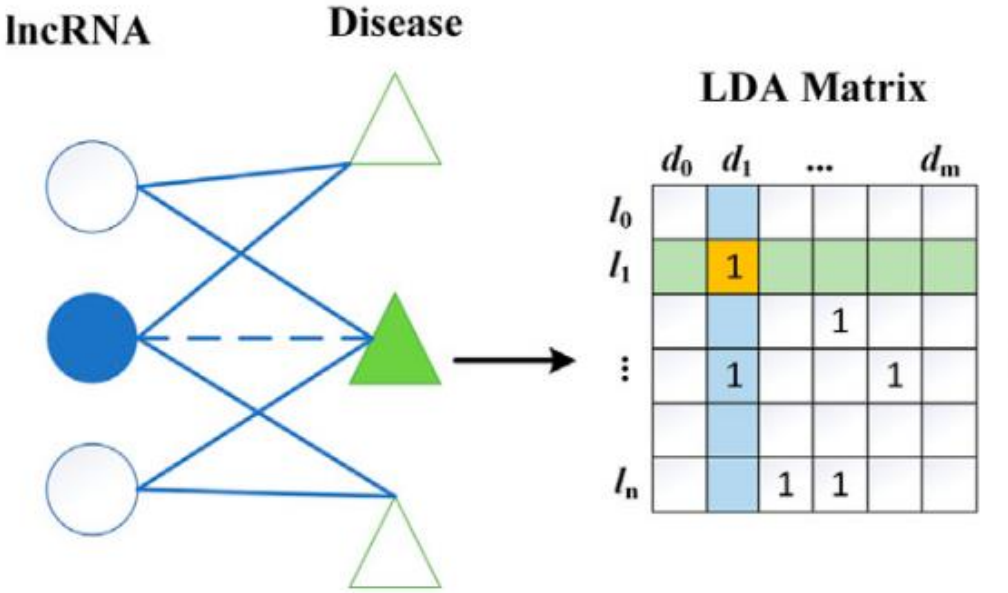
RNAs

- RNAs are divided into:
  - Messenger RNAs
    - Related to protein synthesis
  - Non-coding RNAs
    - Play important role in transcription, translation, epigenetic regulation, immune responses...
    - Multiple types:
      - miRNA: single-strand 21-23 nucleotides
      - siRNA: double-strand 20-24 nucleotides
      - **Long non-coding RNA**: more than 200 nucleotides

LncRNAs

- Recent research reveals that lncRNAs are related to several human **diseases**:
  - Multiple types of cancer:
    - breast, ovarian, lung, gastric, prostate
  - Male infertility
- Wet experiments are still costly, laborious and not efficient

Predicting lncRNA Disease Association



Predicting lncRNA Disease Association

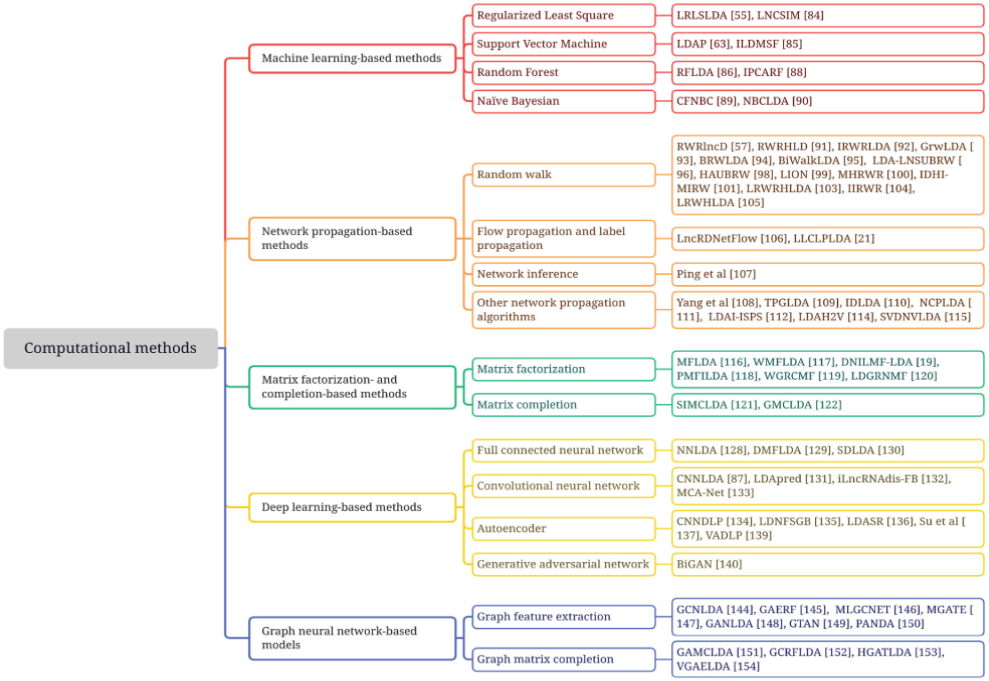
- lncRNAs

ID	Sequence
NR_075089	AAAAATGTTGAATATCAACAATGGGTT....
NR_038832	AAAACCCTATTCATTTGGCACCCAAATGTCATATAGCCGGAACTGGG....
NR_047528	AAAACTGTAATACTCAGACAGACAATTAAGCTGTAAATGAAGTTTTCTTTATGATGC....
AK042766	AAAAGCAGGATGCTTAGAGGGATTGAGTTTGACAAAAGAGAAGGTGGCT....

- Diseases

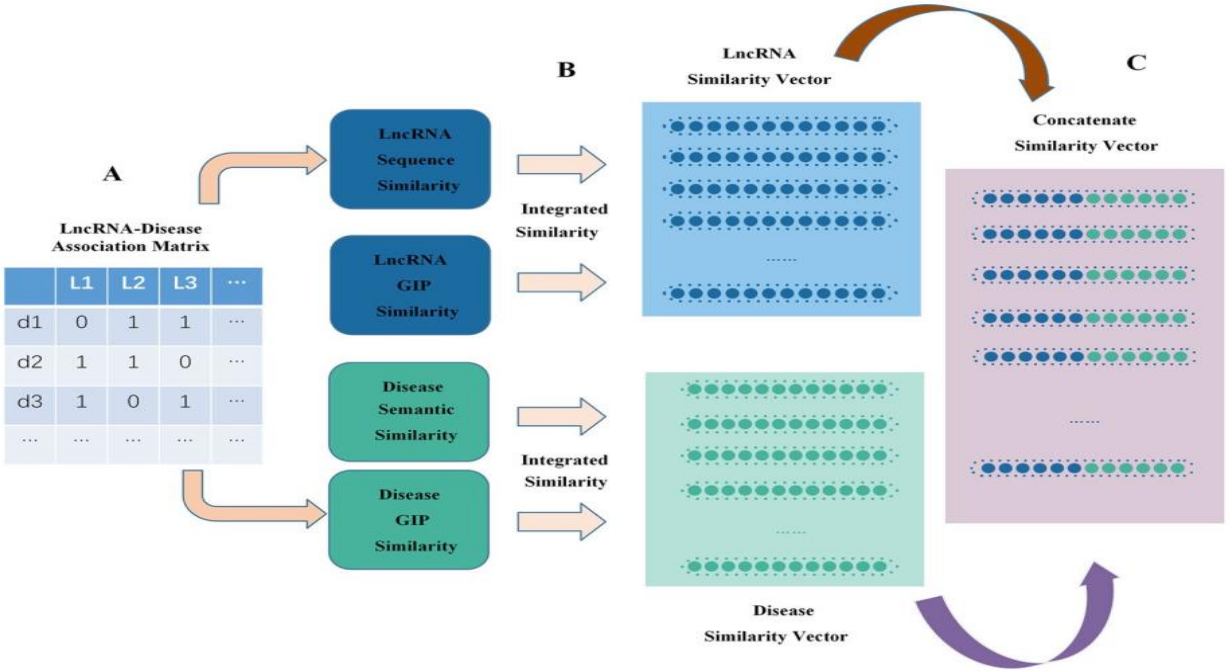
ID	Diseases
NR_075089	Carcinoma, Alzheimers, Heart Failure,….
NR_038832	Melanoma, Multiple Sclerosis,...
NR_047528	Carcinoma, Lung Neoplasms, ….
AK042766	Parkinson, Melanoma, ….

Literature overview



Sheng, Nan, et al. "Data resources and computational methods for lncRNA-disease association prediction." *Computers in Biology and Medicine* 153 (**2023**): 106527.

Most used approach = binary transformation

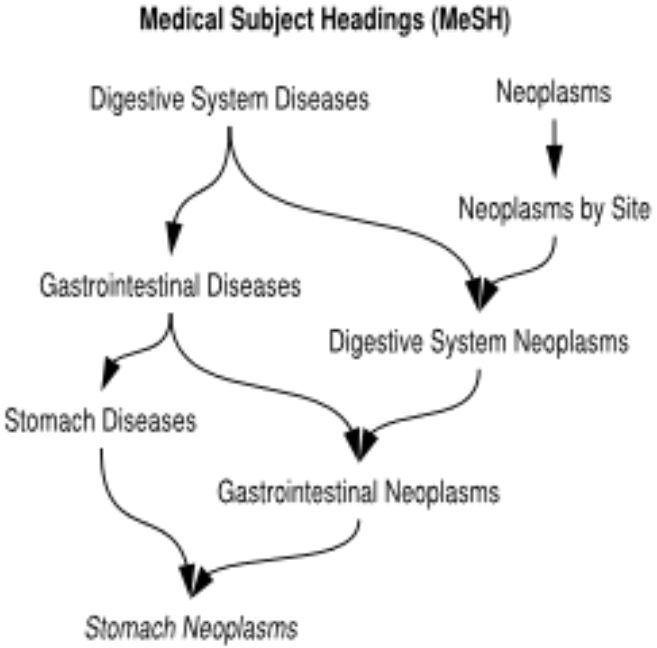


Feature extraction - Diseases

- Build a symmetric matrix in the shape |D| x |D|, |D| = number of diseases
  - Each cell corresponds to the similarity between D_i and D_j
  - General method:
    - Create the cartesian product of D = {(D_1, D_1), (D_1,D_2), (D_1,D_3)., (D_n, D_n)}
    - Compute similarity of each item pair in the cartesian product
  - Survey: Cheng, Liang, et al. "Computational methods for identifying similar diseases." *Molecular Therapy-Nucleic Acids* 18 (2019): 590-604.
    - The most used method is Wang's similarity :
      - Requires a hierarchy of diseases: MeSH

MeSH: MEdical Subject Headings

- "The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. "



Wang's similarity

- Given a directed acyclic graph (DAG), MeSH
- Follow 2 steps for each pair of diseases (d_i and d_j):
  1. Compute the semantic contribution of the diseases on the path to the root (S_{di} (d_j)) to d_i

$$S_{d_i}(d_j) = \begin{cases} 1 & if\ d_j = d_i \\ \max\left\{\gamma * S_{d_i}\left(d'_j\right) | d'_j \in children\ of\ d_i\right\} & if\ d_j \neq d_i \end{cases} \quad (3)$$

Wang's similarity

2. Sum the semantic contribution of each disease (node) on the path from d_i to the root

$$DV_{d_i} = \sum_{d \in T(d_i)} S_{d_i}(d)$$

3. Repeat steps 1 and 2, for disease d_j

4. Wang's Similarity, DS (d_i, d_j): sum of semantic contributions of the intersection of diseases between d_i and dj, divided by DV_{di}

$$DS(d_i, d_j) = \frac{\sum_{d \in T(d_i) \cap T(d_j)} (S_{d_i}(d) + S_{d_j}(d))}{DV_{d_i} + DV_{d_j}}$$

Feature extraction - lncRNAs

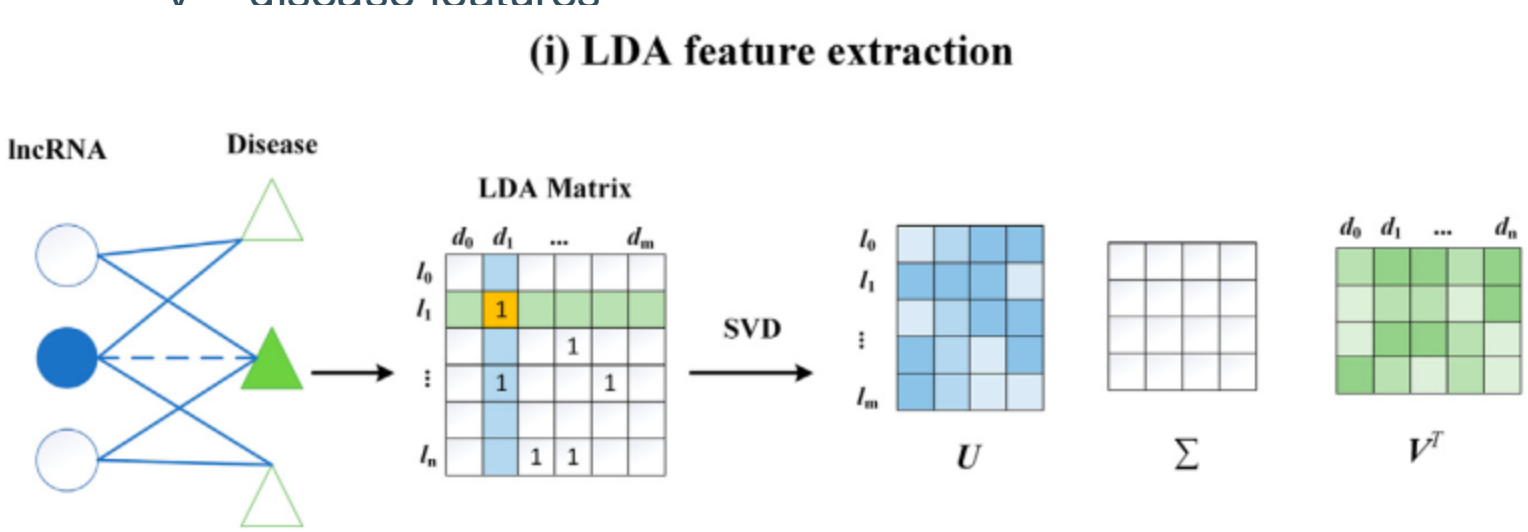
- Similarities:
  - Sequence similarity: Levenshtein/edit distance
    - "the minimum cost to convert a string (RNA) to another string (the other RNA)"
    - Done using insertion, deletion and replacement. Each operation has a cost
  - Functional similarity: Gaussian kernel similarity
    - Given 2 lncRNAs, l_1 and l_2
    - A = association matrix (labels)

$$GKL(l_i, l_j) = \exp\left(-\lambda ||A(l_i) - A(l_j)||^2\right)$$

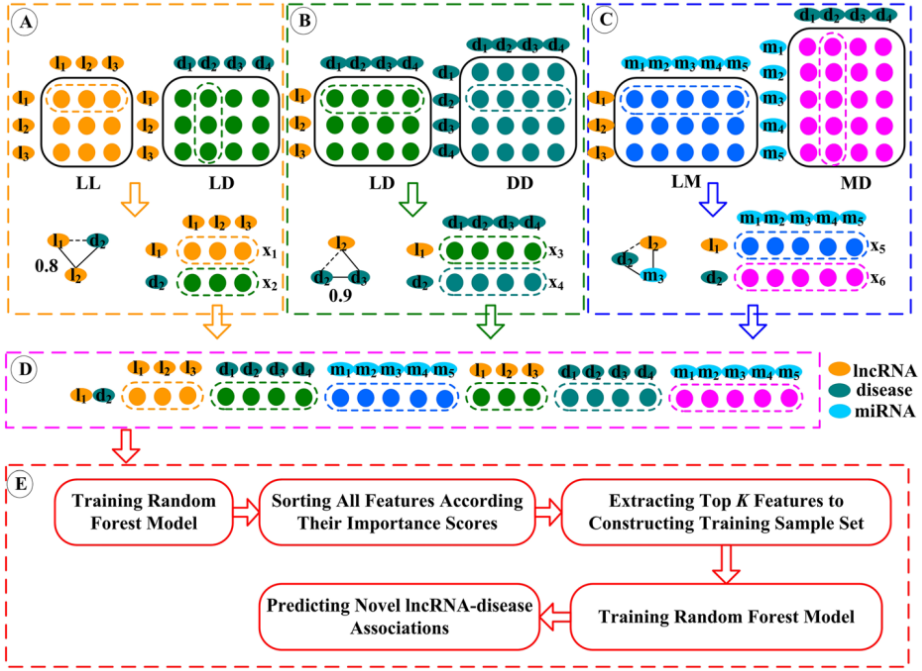
    - Also applicable for diseases. Transpose A.

Feature extraction – RNA and disease

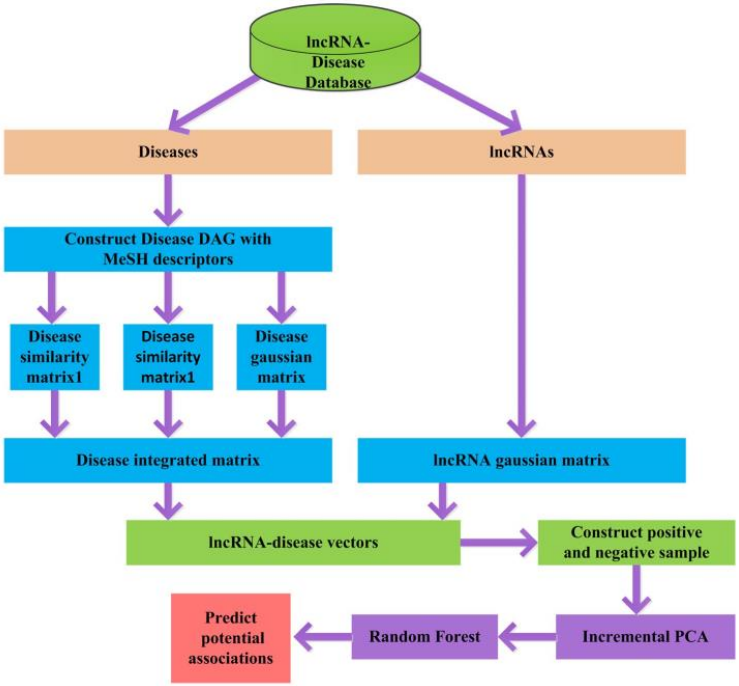
- SVD:
  - U = RNA features
  - V = disease features



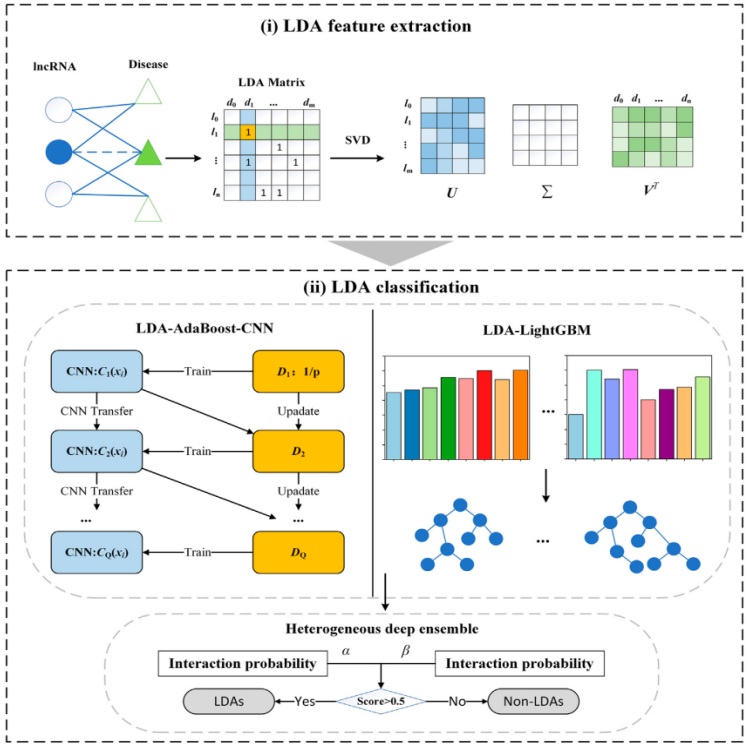
Methods – RFLDA (2020)



Methods – IPCARF (2021)



Methods – LDA SABC (2024)



Considerations

- Innovation relies on novel ways of extracting features:
  - Different kinds of similarities, SVD, etc
- "All" works are binary and (semi) **transductive?:**
  - RNA feature extraction relies on the interaction matrix
  - Features from RNAs could be inductive:
    - Sequence alignment: BLASTn
    - K-mers: Number of occurrences of substrings of size K, e.g., K=2, count AA, AU, AG, AC, UA,…. CC
- *"Inconsistent"* evaluation setup
  - Different cross-validation setups
  - Newly proposed methods do not necessarily compare against each other: state-of-the-art not clearly defined

Our proposal

Multi-label classification

- Use the matrix association directly as output
  - No binary conversion
  - Extract features from the RNAS

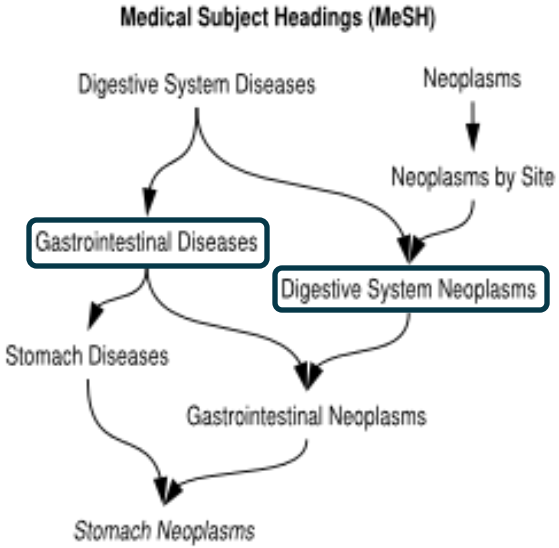
	f_1	f_2	f_3	d_1	d_2	d_3	d_n
1	0	1	34	0	1	1	...
2	2	2	4	1	0	1	...

- Advantages:
  - Huge literature on multi/weak-label classification
    - AFAIK, the uncertainty on negative annotations are ignored
  - Directly exploit label (disease) correlation while building the model
    - Binary approach exploits it in the feature space: disease similarity

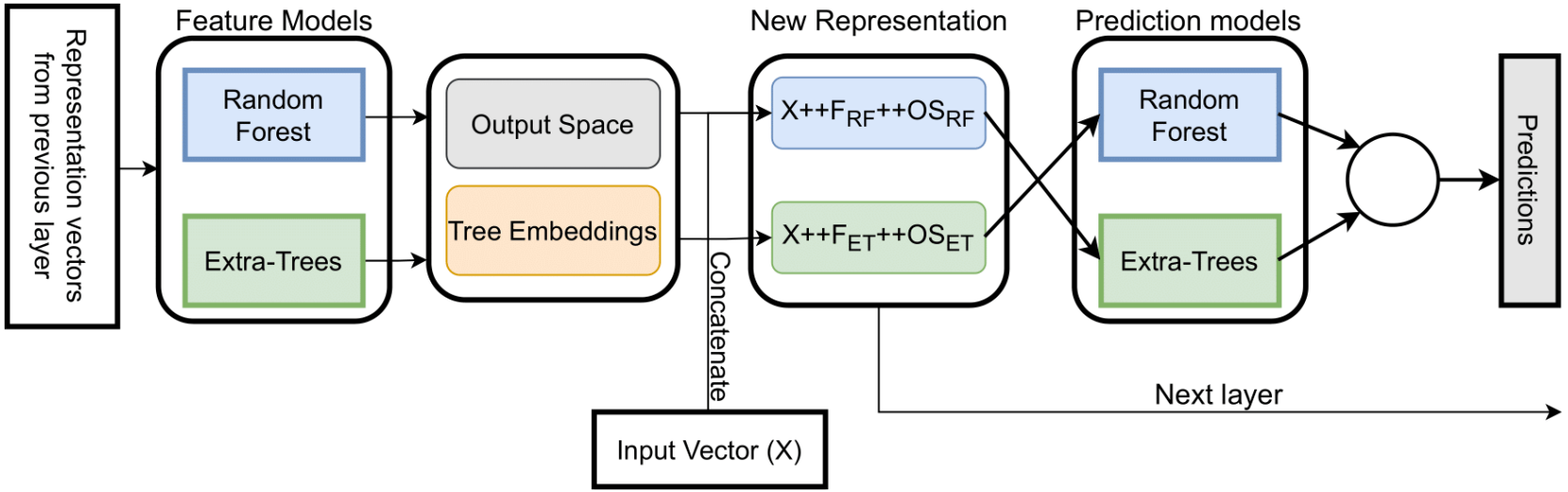
Hierarchical Multi-label classification

Incorporate hierarchical dependencies from MeSH

- Each disease is converted to its path to the root
- Stomach diseases becomes:
  - Gastrointestinal diseases and Digestive System Diseases
- Digestive System Neoplasms becomes:
  - Digestive System Neoplasms, Digestive System Diseases, Neoplasms by Site and Neoplasms



Our proposal – Deep tree-ensembles



Inductive experimental setup

Materials

- Dataset: LncRNADisease version 2.0
  - Publicly available for download
  - Filter to only include diseases with more than 5 annotations

#LncRNAs	#Diseases	#Flat Annotations
355	44	1132/10175 (5%)

- Convert to hierarchical using MeSh

#LncRNAs	#Diseases	#Hierarchical Annotations
355	369	9539/130995 (7%)

Comparison methods

- Binary and (hierarchical) multi-label versions
- Baselines:
  - RFLDA (2020): random forest with top features according to feature importance;
  - IPCARF (2021): incremental PCA + random forest;
  - RF + ET = Random forest + extra trees
- Competitors:
  - Deep-Forest:
    - GcForest: original deep forest, predictions are used as extra features
    - CAFÉ (2022): our deep forest method, tree-embeddings
    - CAFE-FLA (2024): our CAFÉ method with label imputation
  - LDA-SABC (2024): Adaboost-CNN + lightGBM: only binary

Features

- LncRNAs: K-mers with K = ?
  - K-mers: Number of occurrences of substrings of size K, e.g., K=2, count AA, AU, AG, AC, UA,…. CC
  - Total number of features = **827**
- Disease: Wang's similarity – only used in *binary classification*
  - Range 0 to 1
  - Symmetric matrix **44 x 44**

Results 10 folds CV (b) binary, (mc) multilabel

	Hamming error	Label ranking error	Micro ROC	Micro AUPRC	Precision	Recall	F-score	Accuracy
IPCARF (b)	0.109	0.38	0.725	0.273	0.278	0.69	0.384	0.18
RFLDA (b)	0.106	**0.252**	0.805	0.329	0.287	0.778	0.416	0.157
CAFÉ (b)	**0.045**	0.721	0.614	0.17	**0.589**	0.327	0.392	**0.183**
RF + ET (b)	0.119	0.383	0.759	0.312	0.251	0.708	0.366	0.116
GcForest (b)	0.121	0.393	0.758	0.279	0.255	0.719	0.37	0.161
LDA-SABC (b)	0.084	0.267	0.689	0.284	0.326	0.593	0.406	0.186
CAFE (mc)	0.105	0.303	0.791	0.361	0.292	0.752	0.414	0.085
CAFE-fla (mc)	0.119	0.333	0.78	0.288	0.26	0.734	0.379	0.088
CAFÉ (hmc)	0.096	0.266	**0.81**	**0.403**	0.314	**0.779**	**0.444**	0.107
CAFE-fla (hmc)	0.136	0.392	0.734	0.231	0.225	0.689	0.335	0.014

Binary vs multilabel results 10 folds CV

	Hamming error	Label ranking error	Micro ROC	Micro AUPRC	Precision	Recall	F-score	Accuracy
IPCARF (b)	0.109	0.38	0.725	0.273	0.278	0.69	0.384	0.18
IPCARF (mc)	0.084	0.245	**0.823**	**0.44**	0.356	0.773	**0.483**	0.141
RFLDA (b)	0.106	0.252	0.805	0.329	0.287	**0.778**	0.416	0.157
RFLDA (mc)	0.083	**0.239**	0.809	0.409	0.339	0.748	0.464	0.166
GcForest (b)	0.121	0.393	0.758	0.279	0.255	0.719	0.37	**0.161**
GcForest (mc)	0.112	0.337	0.776	0.335	0.278	0.721	0.392	0.057
RF + ET (b)	0.119	0.383	0.759	0.312	0.251	0.708	0.366	0.116
RF + ET (mc)	0.091	0.266	0.806	0.364	0.329	0.774	0.457	0.132
CAFÉ (b)	**0.045**	0.721	0.614	0.17	**0.589**	0.327	0.392	0.183
CAFE (mc)	0.105	0.303	0.791	0.361	0.292	0.752	0.414	0.085

Multilabel vs HMC results 10 folds CV

	Hamming error	Label ranking error	Micro ROC	Micro AUPRC	Precision	Recall	F-score	Accuracy
IPCARF (mc)	0.084	0.245	0.823	**0.44**	0.356	0.773	0.483	0.141
IPCARF (hmc)	**0.083**	**0.236**	**0.825**	0.436	**0.357**	**0.782**	**0.484**	0.181
RFLDA (mc)	**0.083**	0.239	0.809	0.409	0.339	0.748	0.464	0.166
RFLDA (hmc)	0.086	0.246	0.809	0.421	0.333	0.748	0.459	0.118
GcForest (mc)	0.112	0.337	0.776	0.335	0.278	0.721	0.392	0.057
GcForest (hmc)	0.126	0.383	0.709	0.289	0.237	0.684	0.347	0.028
RF + ET (mc)	0.091	0.266	0.806	0.364	0.329	0.774	0.457	0.132
RF + ET (hmc)	0.097	0.272	0.801	0.354	0.317	0.773	0.444	0.071
CAFE (mc)	0.105	0.303	0.791	0.361	0.292	0.752	0.414	0.085
CAFÉ (hmc)	0.096	0.266	0.81	0.403	0.314	0.779	0.444	0.107

Considerations

- Using the (hierarchical) multi-label approach improves current methods from the literature
  - Specially IPCARF and RFLDA
  - The multi-label approach has better computational complexity
  - Incorporating MeSH improve the results in some cases
    - Prune less frequent paths (?)
  - Label imputation does not work yet
    - Sparse annotations (?)

Future work

- Updated dataset LncRNADisease3 (2024):
  - Integrate circRNA-disease associations;
  - 2 hierarchies are available now: Disease Ontology and Medical Subject Headings
- At least 1 more dataset is available: MNDR
- Test other features for the RNAs:
  - Homology-based similarities
  - Sequence embeddings from deep learning methods
- Improve weak-label methods
  - Imputation in sparse datasets

Remarks about the literature

Weird evaluation setup

- Feature extraction is not entirely clear
  - Most of the features relies on information from the interaction matrix (output) -> transductive features
- 5 or 10 fold cross-validation is always reported in the papers
  - AFAIK, masking a percentage of positives (1s) is standard
- Amazing performance in several metrics: more than 90%
  - Report performance on negatives, e.g., recall
    - Recall = TP / (TP + ***FN***)
- Same dataset (different versions): LncRNADisease

RFLDA (2020):

- "The 2697 experiment-supported LDAs are considered as positive samples. The 2697 randomly selected lncRNA-disease pairs not validated by experiments are considered as negative samples. All lncRNA-disease pairs not validated by experiments are taken as unlabelled samples. For 5-fold cross-validation, all positive and negative samples are evenly divided into 5 parts."

with them. Here, we calculate lncRNA functional similarities by Chen et al.'s method [36]. Here, we assume that lncRNA $l(a)$ is related to a group of diseases $DG(a) = \{d(a1), d(a2), \ldots, d(am)\}$, and lncRNA $l(b)$ is related to a group of diseases $DG(b) = \{d(b1), d(b2), \ldots, d(bn)\}$, then the LFS between $l(a)$ and $l(b)$, denoted as $LS(l(a), l(b))$, can be obtained by calculating the similarity between $DG(a)$ and $DG(b)$ by eq. 8.

$$LS(l(a), l(b)) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} (DS(d(ai), d(bj))) + \sum_{j=1}^{n} \max_{1 \le i \le m} (DS(d(bj), d(ai)))}{m + n}$$

IPCARF (2021)

- "Cross-validation is a commonly used method in machine learning that can greatly reduce errors caused by sample selection. In our experiments, we used 10-fold cross-validation (10CV) to assess the classification prediction ability of six different classification algorithms"

Gaussian interaction profile kernel similarity for lncRNA

The Gaussian interaction kernel similarity between lncRNAs L_1 and L_2 is computed as

$$GKS(L_1, L_2) = exp(-k_{lnc}\|L_1 - L_2\|^2), \quad (6)$$

$$k_{lnc} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\|L(i)\|^2}, \quad (7)$$

where n represents the lncRNA number.

LDA-SABC (2024)

- "We partitioned the disease-lncRNA pairs into five equal subsets, employing a four-to-one ratio for training and testing, which facilitated five cross-validation iterations. In each round, we removed all known associations from the test set and evaluated the performance of the trained model on the test samples."
- Upon code inspection:
  1. Select N random positive annotations = positive indexes
  2. Select N random negative annotations = negative indexes
  3. Mask (set to 0) the indexes of steps 1 and 2 in the interaction matrix
  4. Run model on the interaction matrix
  5. Evaluate on indexes from steps 1 and 2
- Evaluation on negatives that were already given as negatives to the model?

LDA-SABC and LDA-VGHB (2024)

- 5-fold CV on independent lncRNAs and independent diseases: First, 20% of lncRNAs and 20% of diseases were randomly selected to construct a ‘node test set’. Next, the remaining lncR-NAs and diseases were taken as a ‘node train set’. Third, all edges linking a node in the ‘node train set’ with a node in the ‘node test set’ were removed. Finally, one learner was trained only on the‘node train set’ to find potential LDAs within the ‘node test set’
- Upon code inspection:
  - Apply SVD on the unmasked association matrix
    - Extracting RNA and disease features
  - Perform cross-validation on the unmasked association matrix
  - Data leakage (?): interactions are already present in the SVD