Prediction of miRNA-disease associations with vector space model

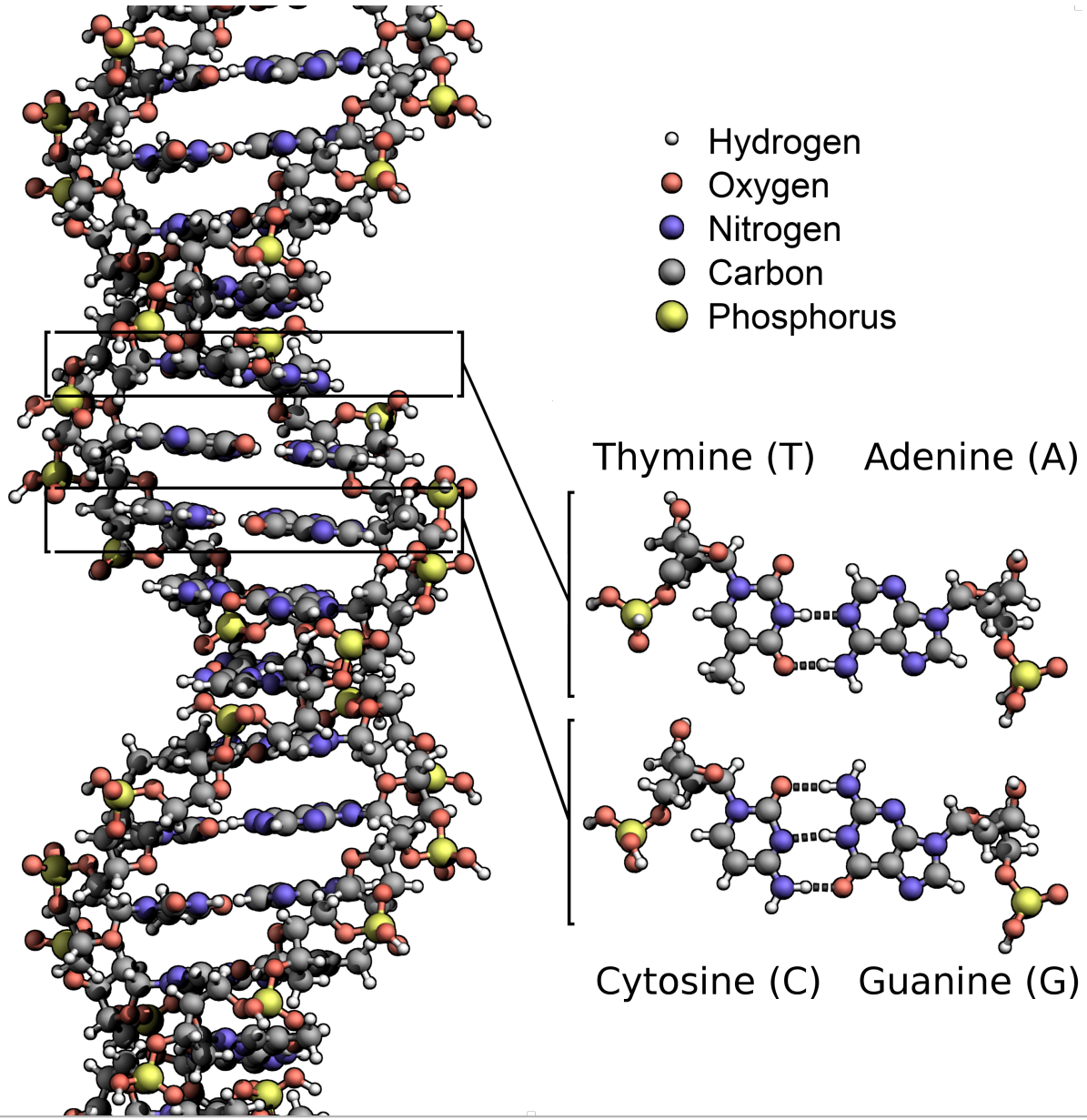
Claude Pasquier[1]

December 4, 2015

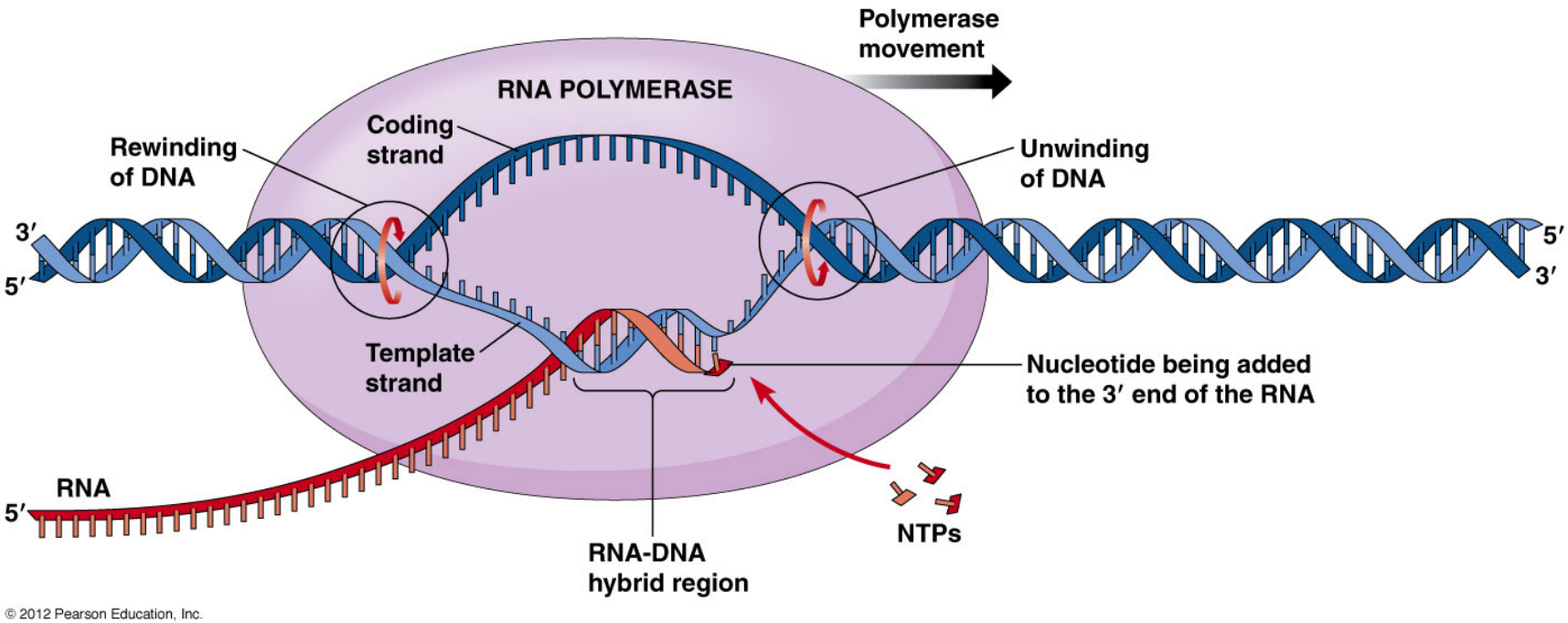
[1]Team SPARKS - I3S - Sophia Antipolis

claude.pasquier@unice.fr

DNA (Deoxyribonucleic acid)

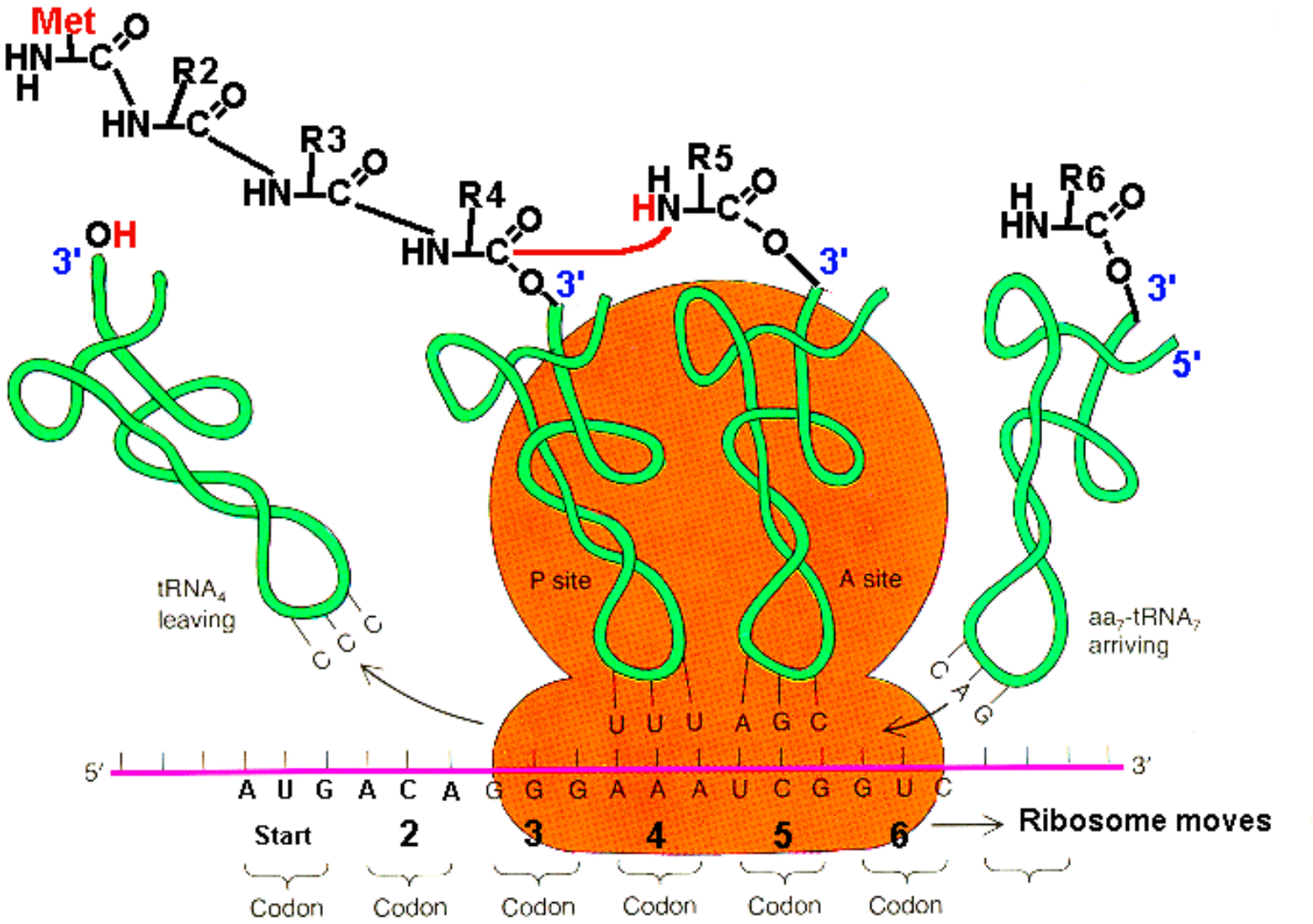


Transcription



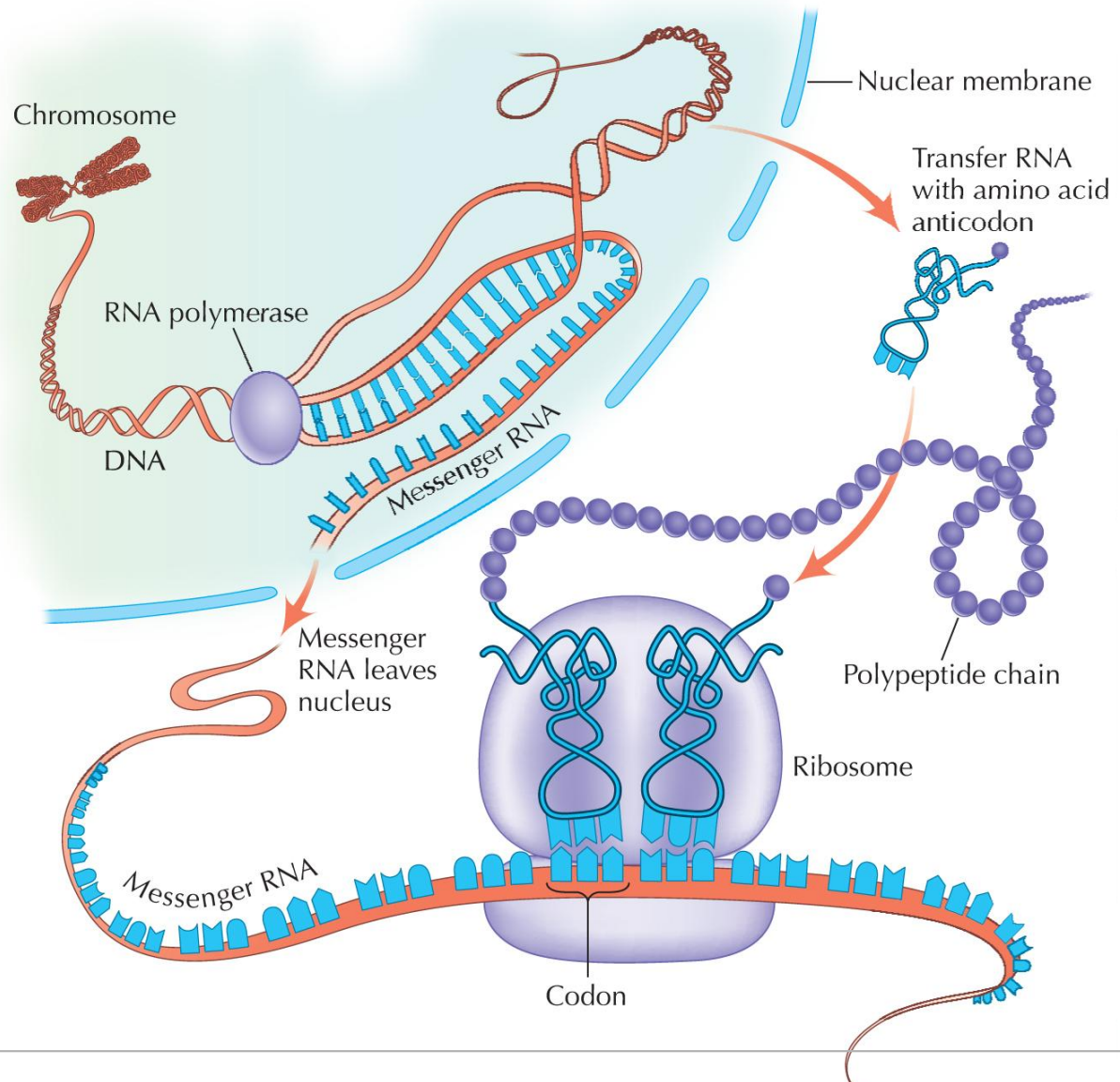
In RNA, Tymine (T) is replaced by Uracyl (U)
A RNA is composed of a sequence of A, C, U, G

Translation

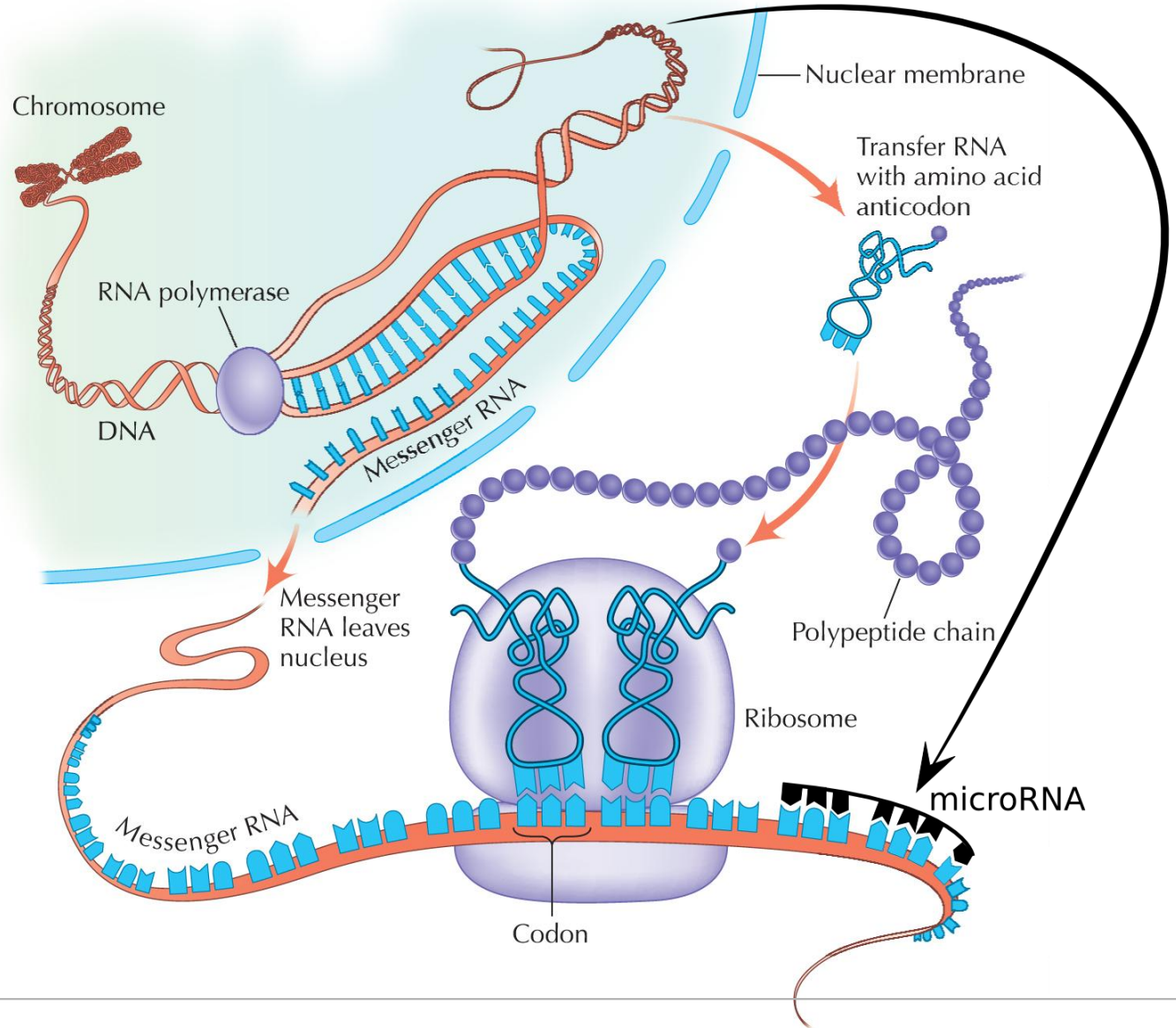


Modified from Griffiths et al., AN INTRODUCTION TO GENETIC ANALYSIS, 6th Ed., W.H. Freeman & Co., 1996.

Protein synthesis



miRNA action



Objective

To increase our knowledge
of miRNA-disease associations
by using computational methods

Motivation

- MicroRNAs play critical roles in many physiological processes.
- Their dysregulations are also closely related to the development and progression of various human diseases, including cancer
- Therefore, identifying new microRNAs associated with diseases contributes to a better understanding of pathogenicity mechanisms.

Related work

Methods based on disease similarities (2008)

limited performances

Methods based on target similarities (2010-2013)

pb: as the number of verified miRNA-mRNA interactions is currently low, predicted interactions (with high false-positive and false-negative rates) are also used

Methods using supervised learning (2013)

pb: the methods need negative miRNA-disease associations

Methods based on miRNA functional similarities (2012-2015)

pb: miRNA functional similarities remain largely unknown. All the methods use predicted scores

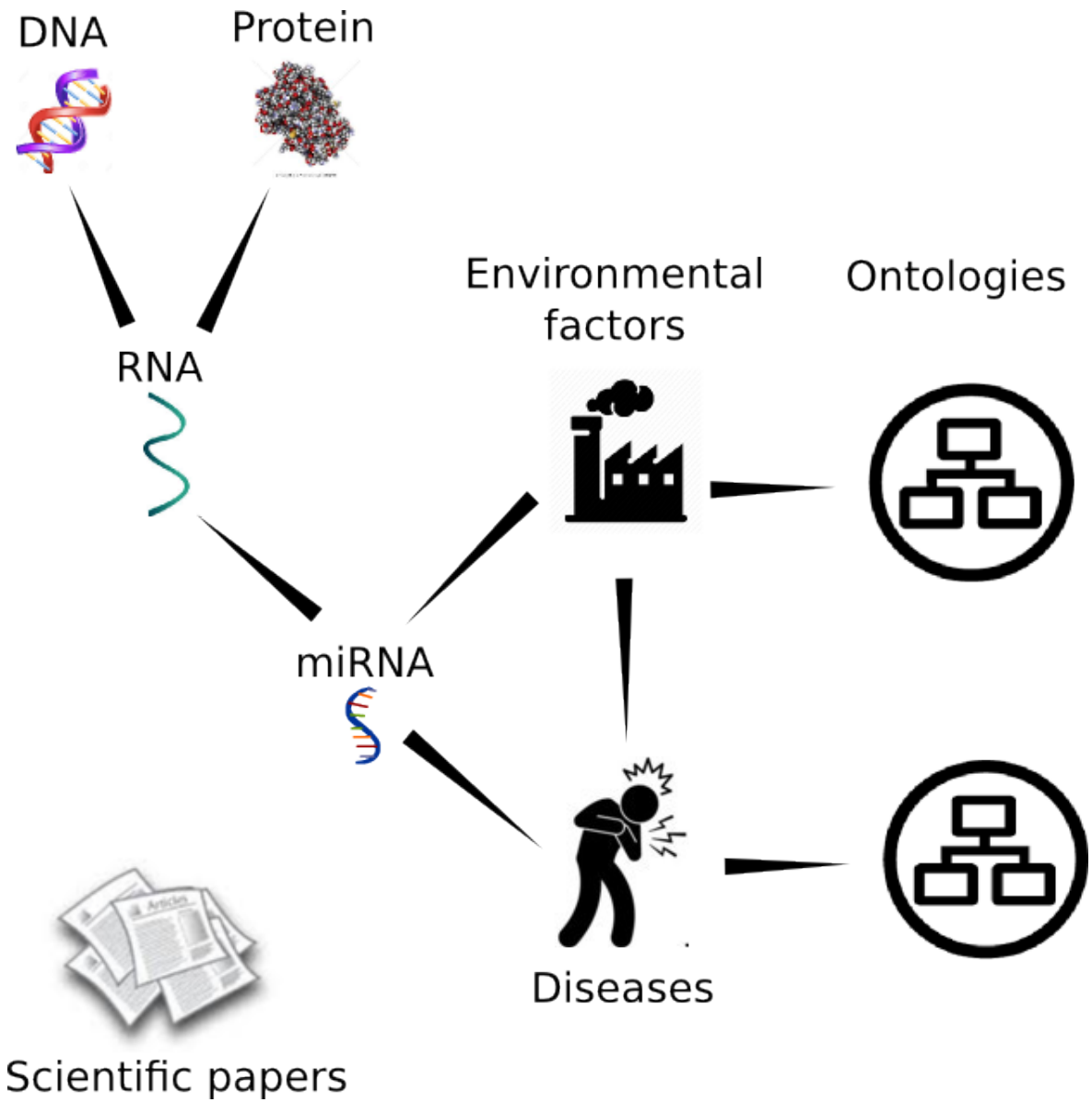
methods using several sources of data (2012-2015)

Best performances obtained with these methods

Observations

1. best performance is reported with methods that use different sources of data
2. some knowledge available in the domain is not used in existing programs
   - miRNA families suggest a common structure configuration in sets of genes that hint to shared functions,
   - significant correlation between pairs of expressed miRNAs occurs at a distance < 50 kb, suggesting common regulatory pathway,
   - important information about miRNAs is contained in scientific papers.
3. each miRNA is associated with a set of heterogeneous data that makes difficult the use of traditional machine learning algorithms.

Data availability

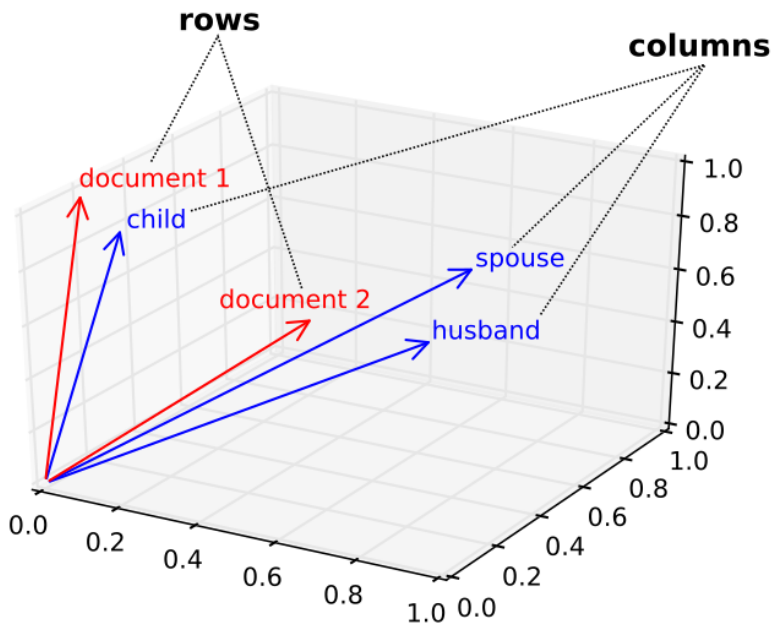


Idea: using distributional semantics

- Distributional semantics
  - based on an old hypothesis in linguistics (Zellig Harris,1954) that states that the meaning of words can be deduced from their distributions (context in which the words are used)
- Computational systems proposed from late 80'
  - spacial models (vector space model)
  - e.g. Latent Semantic Analysis (LSA), US PATENT in 1988
  - Latent = identification of hidden relationships
- Used successfully in various applications
  - finding semantic similarity between words
  - word sense disambiguation
  - information retrieval (search engines)

Vector space modeling

- Collecting distributional information in high-dimensional vectors
  - each "item" is represented by thousands or millions of values
- Performing dimensionality reduction of the original space
  - use of Singular Value Decomposition
    (similar to Principal Component Analysis)
  - this keeps most important semantic information
  - this reduces noise and other undesirable artifacts
- Defining semantic similarity in terms of vector similarity



Our hypothesis

If we consider that

- "items" represent miRNAs
- data associated with miRNAs are stored in vectors of values

Then

new knowledge about miRNAs might be highlighted
by using distributional semantics

Problem

Vector Space Model is designed to process textual data

Textual data

For textual data, the 'classic' following steps can be performed

1. collecting scientific papers associated with miRNA
2. constructing a miRNA-term document matrix
3. weighting miRNA-term relationships (e.g. use of TF/IDF)
4. mapping data to a lower-dimensional space
5. exploring the reduced space to find miRNA-term associations

Non textual data

Non textual data represent biological facts

- Links between miRNAs and diseases, targets or families consist of binary information - There is no weight - This is True or False
  - Some author weight each association by counting the number of time the link appears in databases. This is not a good idea (selection bias)
- Neighboring between miRNAs are expressed as integers representing a distance

We need to define several weighting schemes

Weighting miRNA-disease associations

Weight of an association between a miRNA m and a disease d is obtained with:

$$x_{md} = max(simil(d, k)) \ for \ k \ in \ \{w | x_{m,w} > 0\}$$

where *simil* is a similarity measure taking into account the distance between two diseases in the MeSH ontology.

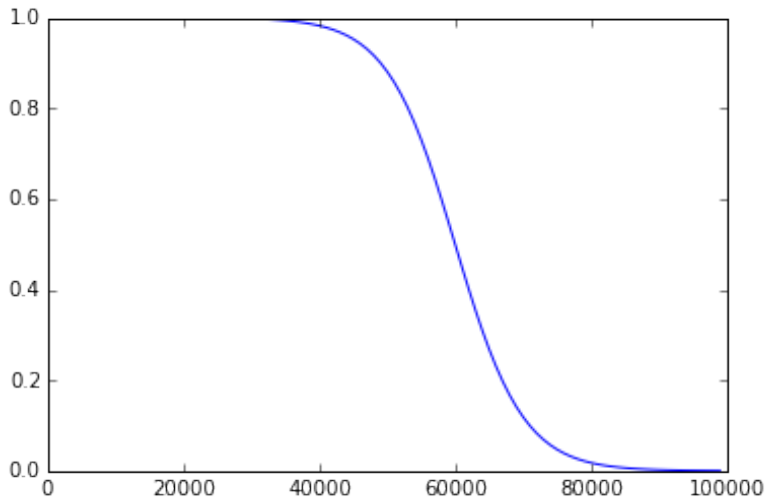
Weight between miRNA m and disease d is the maximum similarity between d and every disease associated with m

Weighting miRNA-neighbor associations

Genomic distance is transformed to a weight, expressing the correlation between two miRNA, with a sigmoid function of the form:

$$w(dist) = 1 - \frac{1}{1 + e^{-k(dist - dist_0)}}$$

where *dist* is the genomic distance between two miRNAs, $dist_0$, the sigmoid midpoint, is set to 6×10^5 (a distance of 60kb is associated with a correlation of 0.5) and k, the steepness of the curve, is set to 2×10^{-4}.



Weighting miRNA-target associations

Associations can be represented by a bipartite network connecting items from a set of miRNAs to items from a set of targets

On this network, we use a recommender system that was primarily designed for online shops to predict user rating or preferences.

Network Based Inference (Zhou et al. 2007

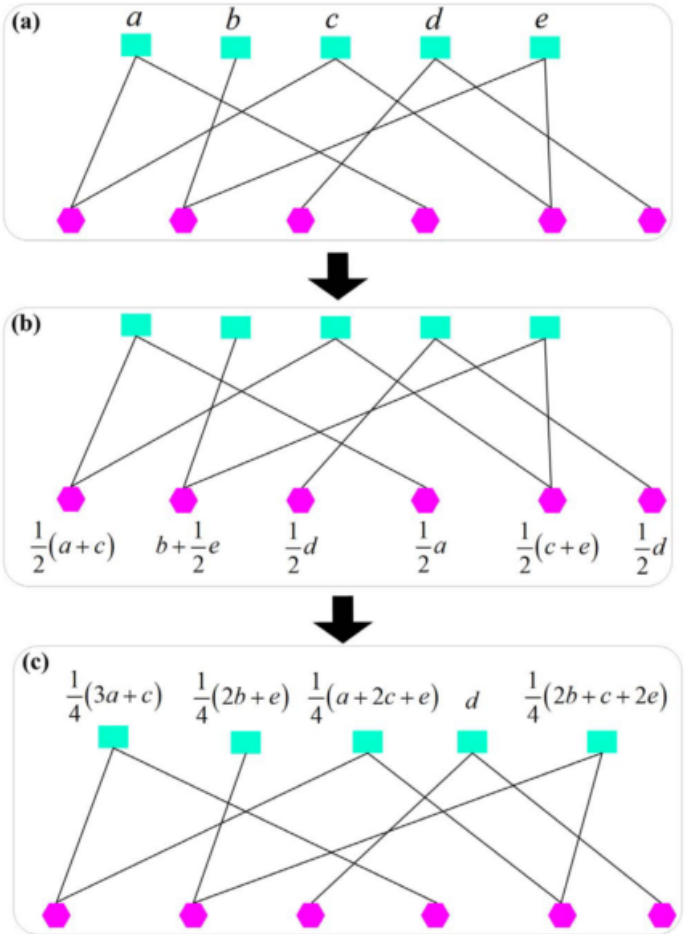


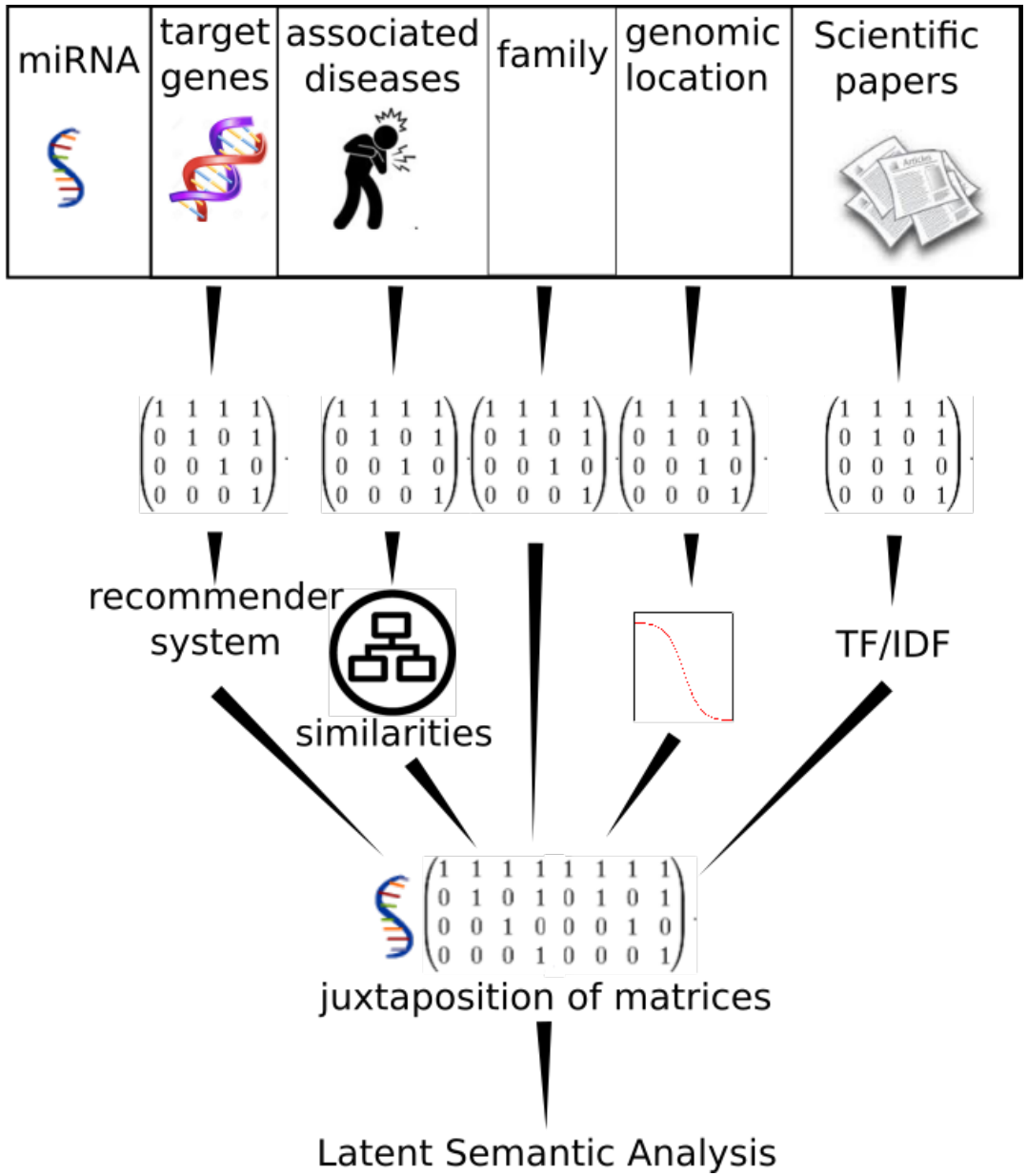
Figure 3. Principle of the resource-allocation process in a bipartite network. The green rectangles represent X nodes and red hexagons represent Y nodes. The whole process consists of two steps: First, the resource flows from X to Y (a→b), and then returns to X (b→c).
doi:10.1371/journal.pone.0087797.g003

$$M = \begin{pmatrix} a| & 1 & 0 & 0 & 1 & 0 & 0 \\ b| & 0 & 1 & 0 & 0 & 0 & 0 \\ c| & 1 & 0 & 0 & 0 & 1 & 0 \\ d| & 0 & 0 & 1 & 0 & 0 & 1 \\ e| & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

- (c) is represented by a 5x5 matrix (W) representing the information transfer between blue nodes
- The new weighted matrix is obtained by *MxW*

$$\begin{pmatrix} a| & 1 & 0 & 0 & 3/4 & 1/4 & 0 \\ b| & 0 & 3/4 & 0 & 0 & 1/4 & 0 \\ c| & 3/4 & 1/4 & 0 & 1/4 & 3/4 & 0 \\ d| & 0 & 0 & 1 & 0 & 0 & 1 \\ e| & 1/4 & 1 & 0 & 0 & 3/4 & 0 \end{pmatrix}$$

The method

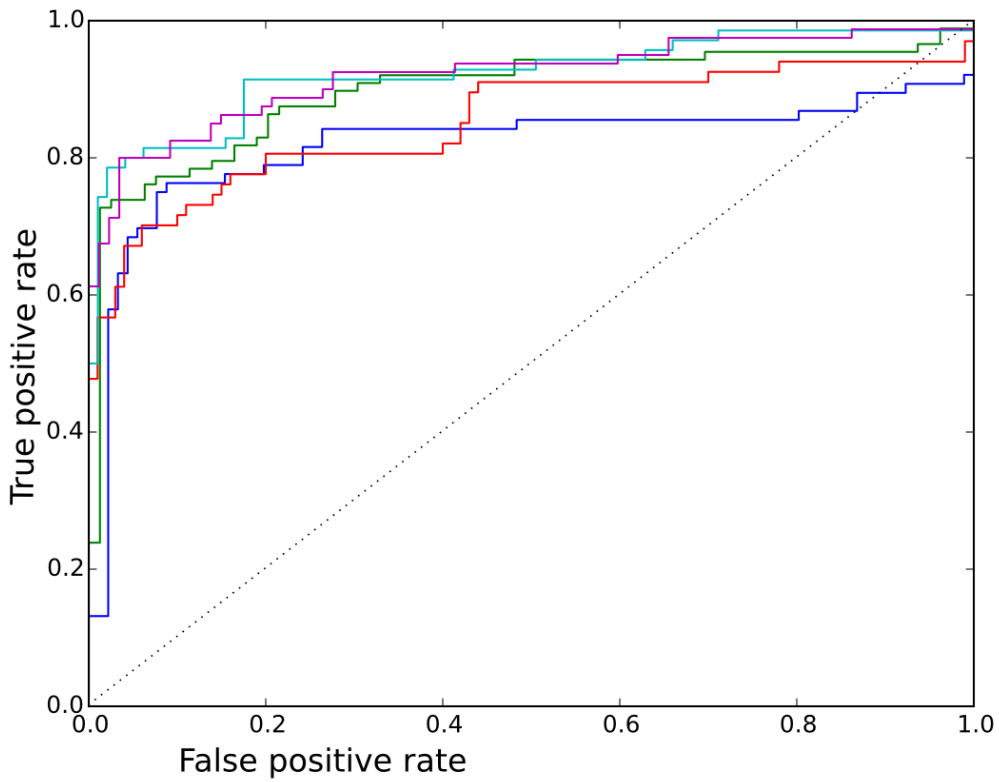

miRNA
target genes
associated diseases
family
genomic location
Scientific papers
recommender system
similarities
TF/IDF
juxtaposition of matrices
Latent Semantic Analysis

Evaluation process

- Based on the experimentally verified miRNA-disease associations stored in the human miRNA-disease database (HMDD) v2.0 of June, 14, 2014
- 5 fold cross validation
  5 partitions: 4/5 of data used for training, 1/5 used for testing
- every partition is used one time for testing
  - with associations to tested disease removed
  - with data collected from litterature also removed
- the predictions are performed for 83 diseases
  (associated with at least 20 miRNAs)
- evaluation consists in measuring whether the program is able to recover the deleted associations

Measuring the performance of the classifier

- plotting Receiver Operating Characteristic (ROC) curve
- an unique measure is obtained with AUC (Area Under the Curve)

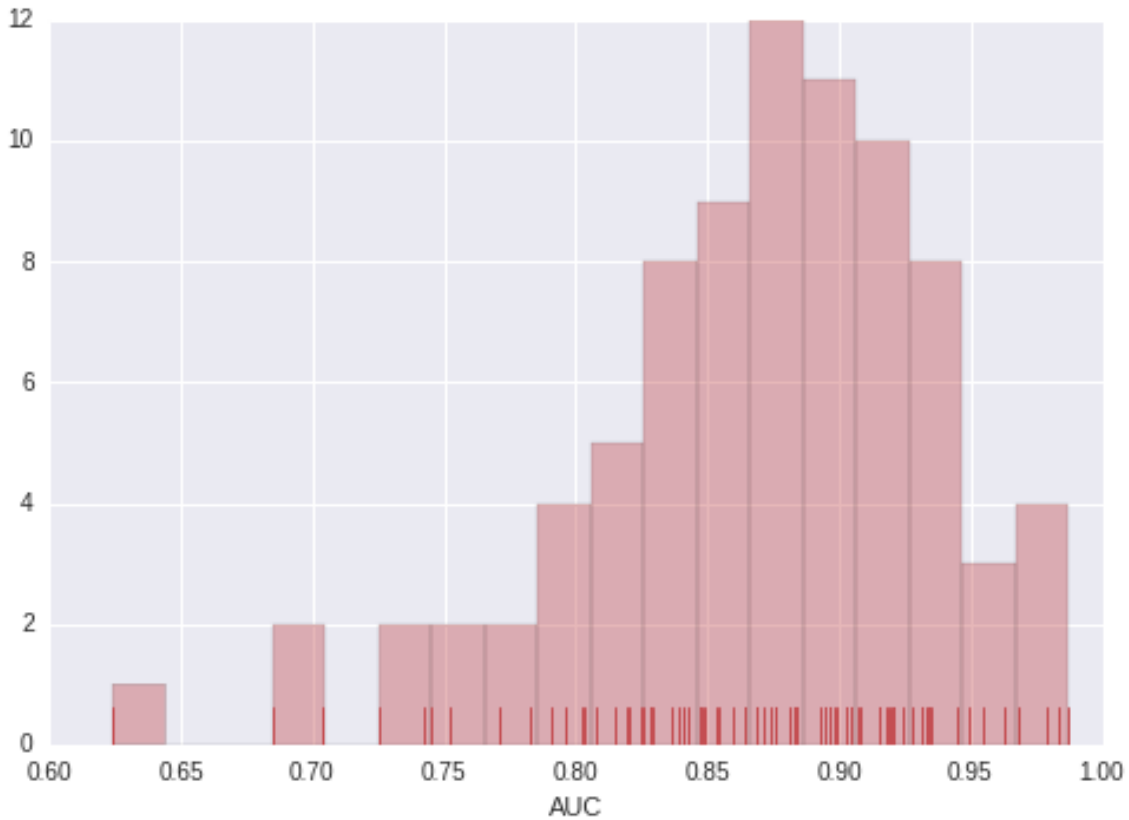


Curves obtained the for the 5 partitions of Breast cancer data

Interpretation of AUC:

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

Univariate distributions of AUC scores



- Average AUC value: 0.867 (min: 0.624, max: 0.987)
- Most of the predictions are above 0.8

Comparison with other methods

Disease name	MIDP 2015	RWR 2012	HDMP 2013	RLS 2014	Chen 2013	MiRAI 2015
Acute myeloid leukemia	0.91	0.84	0.86	0.85	0.72	0.90
Breast neoplasms	0.84	0.79	0.80	0.83	0.65	0.86
Colorectal neoplasms	0.85	0.79	0.80	0.83	0.66	0.86
Glioblastoma	0.80	0.68	0.70	0.71	0.61	0.90
Heart failure	0.82	0.72	0.77	0.74	0.76	0.80
Hepatocellular carcinoma	0.81	0.75	0.76	0.79	0.61	0.81
Lung neoplasms	0.88	0.83	0.84	0.86	0.61	0.90
Melanoma	0.84	0.78	0.79	0.81	0.64	0.85
Ovarian neoplasms	0.92	0.88	0.88	0.91	0.64	0.87
Pancreatic neoplasms	0.95	0.87	0.90	0.89	0.68	0.93
Prostatic neoplasms	0.88	0.82	0.85	0.84	0.63	0.87
Renal cell carcinoma	0.86	0.82	0.83	0.84	0.63	0.87
Squamous cell carcinoma	0.87	0.82	0.82	0.85	0.68	0.88
Stomach neoplasms	0.82	0.78	0.79	0.80	0.63	0.82
Urinary bladder neoplasms	0.90	0.82	0.85	0.85	0.63	0.88
AVERAGE	0.862	0.799	0.816	0.827	0.652	0.867

Prediction results for less studied diseases

Disease name	nb assoc	MiRAI
Arthritis, Rheumatoid	21	0.912
Atrial Fibrillation	21	0.903
Azoospermia	21	0.892
Barrett Esophagus	21	0.966
Coronary Artery Disease	21	0.797
Crohn Disease	21	0.933
Diabetes Mellitus, Type 2	20	0.825
Infertility, Male	20	0.863
Liver Neoplasms	20	0.814
Sarcoma, Kaposi	20	0.920

Analysis of poor results

- Only the following 2 diseases obtain a score lower that 0.70.
  - **Lupus Vulgaris**, associated with 62 miRNAs obtained from an unique paper analyzing the microRNA expression patterns in renal biopsies of lupus nephritis patients.
  - **Adenoviridae Infections** is associated with 74 miRNAs extracted from two tables of an unique paper presenting the up and down-regulated miRNAs in adenovirus type 3 (AD3) infected human laryngeal epithelial.
    - Only 14 associations are really confirmed by Q-PCR.
    - By keeping only verified associations, the AUC increases to 0.796

$\Rightarrow$ The database used to develop the method contains false associations

(manual) analysis of Breast Cancer associations

- Associations extracted from 319 different papers
- 16 of the papers provide more than 10 associations
- 3 papers contain 40 unconfirmed associations
- By removing dubious associations, the AUC score for Breast Cancer jumps from 0.864 to 0.891.

Example of a "problematic" article

Circulating Micro-RNAs as Potential Blood-Based Markers for Early Stage Breast Cancer Detection, Schrauder et al., PLoS ONE 2012.

- microarray-based miRNA profiling on whole blood of 48 early stage breast cancer patients at diagnosis along with 57 healthy individuals as control
- they found 13 significantly up-regulated miRNAs and 46 significantly down-regulated miRNAs that they list in a table
- two miRNAs were validated by RT-qPCR

What next?

Our method achieves state of the art performance on the classification of known association.

$\Rightarrow$ This is good ! But not enough !

The real contribution

Is the method able to predict new (unknown) associations?

Example of prediction obtained for one partition of breast cancer data

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1]

1 = positive
0 = negative

"new" miRNAs associated with breast cancer

miRNA name	described in (PMID)	publication date
hsa-mir-106a	19706389	Sept. 2009
hsa-mir-130a	23528537	Jun. 2013
hsa-mir-138-1	23300839	Dec. 2012
hsa-mir-138-2	23300839	Dec. 2012
hsa-mir-142	25406066	Nov. 2014
hsa-mir-144	25465851	Dec. 2014
hsa-mir-150	24312495	Dec. 2013
hsa-mir-15b	22908280	Sept. 2012
hsa-mir-181c	23524334	Jul. 2013
hsa-mir-19b-2	21059650	Jan. 2011
hsa-mir-208a		
hsa-mir-30e	19432961	May 2009
hsa-mir-378a	20889127	Oct. 2010
hsa-mir-99a	21575166	May 2011

All are confirmed by recent litterature excepted mir-208a that can be proposed as a novel miRNA associated to breast cancer

Recent confirmation as of October 20, 2015

Sun et al.
MiR-208a stimulates the cocktail of SOX2 and β-catenin to inhibit the let-7 induction of self-renewal repression of breast cancer stem cells and formed miR208a/let-7 feedback loop via LIN28 and DICER1.
Oncotarget. 2015 Oct 20;6(32):32944-54

Dubious associations with breast cancer

hsa-mir-1323	hsa-mir-1469	hsa-mir-215	hsa-mir-2355
hsa-mir-3130-1	hsa-mir-3130-2	hsa-mir-3186	hsa-mir-411
hsa-mir-4257	hsa-mir-4306	hsa-mir-718	

- 9 miRNAs come from a table of an unique publication that lists the overexpressed circulating miRNAs. None of them were confirmed
- the 2 remaining miRNAs (mir-215 and mir-411) are specified in an unique paper focusing on circulating miRNAs used as biomarkers

9 out of the 11 association are not verified

According to recent papers, the conclusions based on circulating miRNAs should be taken with caution

Results

list of 811 potential miRNAs associated to 93 diseases

list of 144 potential miRNAs falsely associated with 42 diseases

How to improve accuracy?

- by setting the "optimal" set of parameters to tune the method
  - 37 boolean parameters control the program
  - one run lasts over 4 hours and uses up to 5Gb of RAM
  - $\Rightarrow$ use of genetic algorithms distributed on a grid
- by improving the accuracy of input data
  - biological experiments
  - outliers detection by conceptual clustering
  - graph analysis
  - text mining

Collecting associations with text mining

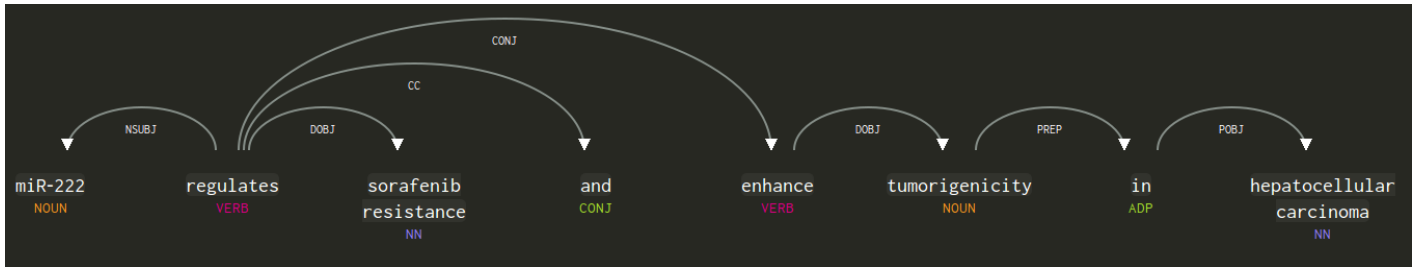
miR-222 regulates sorafenib resistance and enhance tumorigenicity in hepatocellular carcinoma.

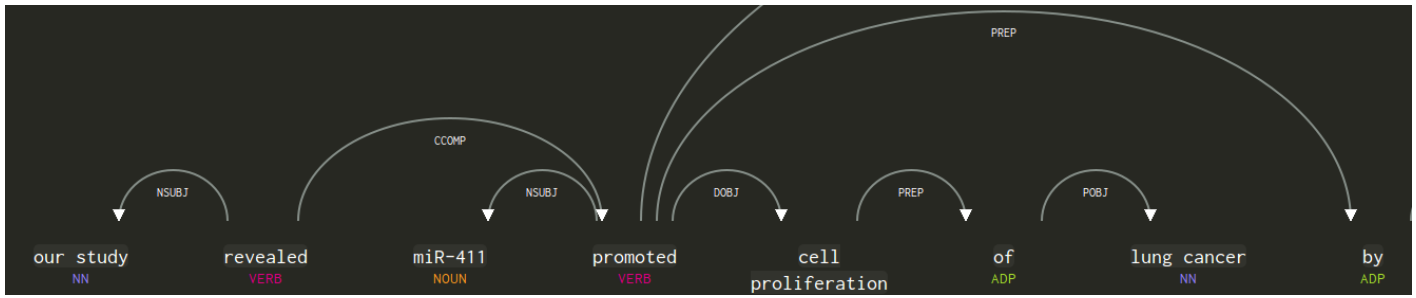
Our study revealed miR-411 promoted cell proliferation of lung cancer by targeting tumor suppressor gene FOXO1.

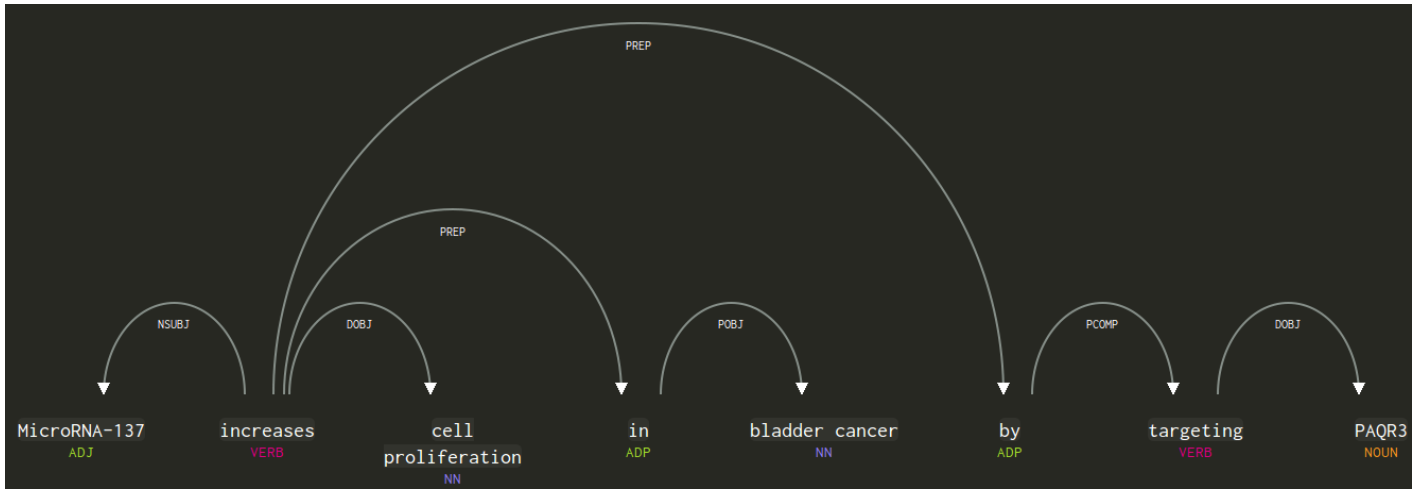
MicroRNA-137 increases cell proliferation in bladder cancer by targeting PAQR3.

Our findings reveal a previously unrecognized aspect of BRCA1's miRNA-mediated post-transcriptional regulation, namely the targeting of its amino acid coding region by the miR-15/107 group of miRNAs

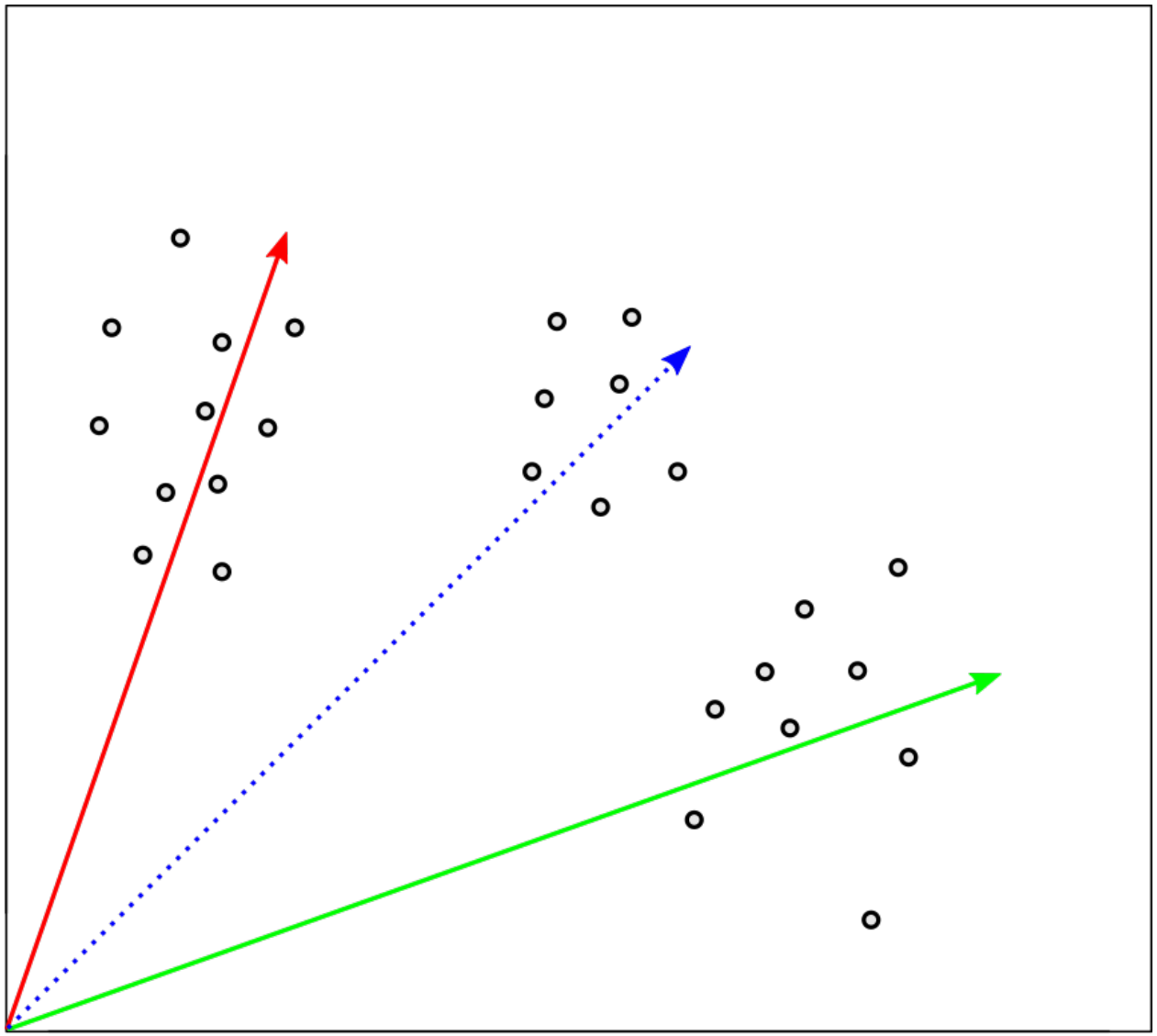
Collecting associations with text mining







Question: can we obtain new knowledge with multi-terms queries?



Question: can we use models that capture more semantics (allowing vector arithmetics)?