Measuring Distances between Medical Entities. Step 1: DrugBank

Alberto Olivares-Alarcos — Iva Stankovic — Humberto González and Horacio Rodríguez
Department of Computer Science, Universitat Politècnica de Catalunya, UPC

Motivation

- **Applicability —** Similarity measurements between entities are essential in several applications and tasks in Artificial Intelligence in general and in Natural Language Processing in particular
- **Challenging —** The problem of having a well stablished numerical distances between semantic entities (drugs, in this case) is still not solved since it's difficulty. On the one hand, there exists a large variety of genres, on the other hand, medical entities have several properties (dimensions) to compute the similarity
- **Scope —** The scope of this work goes farther than computing similarities between drugs. Our aim is to do the same for other medical entities (e.g. anatomical parts, diseases, etc.)

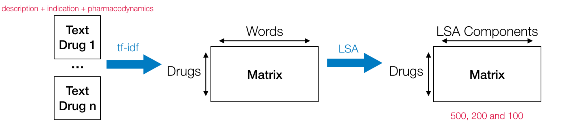
Overview

- **Data —** All data used along this work is extracted from DrugBank (version 5.0.11, released 2017-12-20). The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information
- **Implementation —** Three different similarity measures are computed, using different properties or dimensions of the drug data: textual, taxonomic (both semantics) and molecular information
- **Evaluation —** The computed similarities are evaluated indirectly (clustering based) and directly (ground truth based)

IMPLEMENTATION

TEXTUAL SIMILARITY

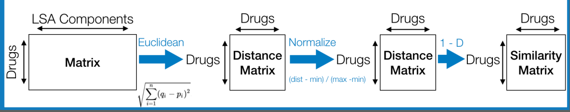
Data Representation


description + indication + pharmacodynamics

Text Drug 1 … Text Drug n — tf-idf → Drugs × Words Matrix — LSA → Drugs × LSA Components Matrix

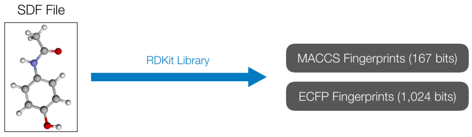
500, 200 and 100

Similarity Measure


Drugs × LSA Components Matrix — Euclidean $\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$ → Drugs × Drugs Distance Matrix — Normalize (dist - min) / (max -min) → Drugs × Drugs Distance Matrix — 1 - D → Drugs × Drugs Similarity Matrix

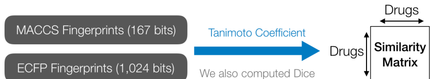
MOLECULAR SIMILARITY

Data Representation


SDF File — RDKit Library →

- MACCS Fingerprints (167 bits)
- ECFP Fingerprints (1,024 bits)

Similarity Measure


MACCS Fingerprints (167 bits) / ECFP Fingerprints (1,024 bits) — Tanimoto Coefficient → Drugs × Drugs Similarity Matrix

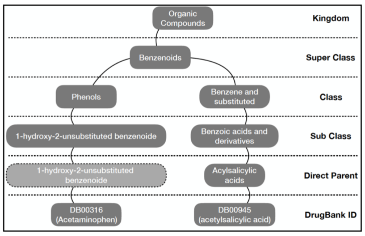
We also computed Dice Coefficient but it is equivalent to Tc.

Jaccard (Tanimoto). $S(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$

TAXONOMIC SIMILARITY

Data Representation

Sub-graph Example

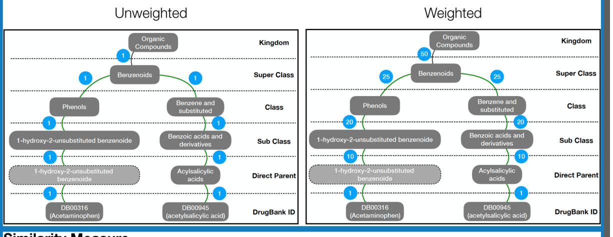


Number of drugs	1,661
Number of nodes	2,360
Number of edges	2,452
Average degree*	2.0780
Directed**	No
Radius	7
Center	Organic acids and derivatives
Density***	0.0008
Depth **(of each tree)**	5

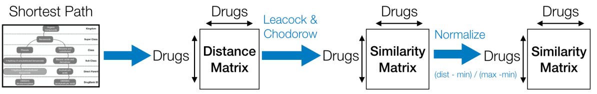
* Average number of edges of adjacent to each node
** Edges with direction
*** Density of edges over nodes

Shortest Path Computation

Unweighted — Weighted



Similarity Measure

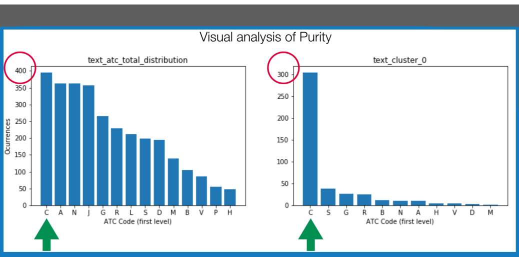

Shortest Path → Drugs × Drugs Distance Matrix — Leacock & Chodorow → Drugs × Drugs Similarity Matrix — Normalize (dist - min) / (max -min) → Drugs × Drugs Similarity Matrix

Leacock & Chodorow — **sim(d1,d2) = -log(distance / (2 · depth))**

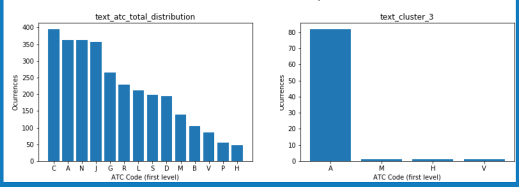
EVALUATION AND RESULTS

CLUSTERING BASED EVALUATION (TEXTUAL)

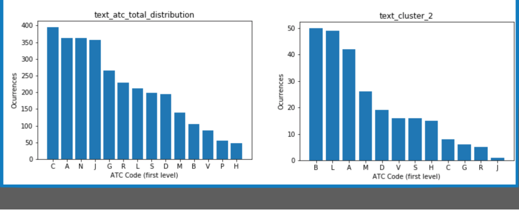
Visual analysis of Purity



Most common ATC Code is predominant



Difficult to extract a conclusion



SETUP

Experiment	Number of Drugs	ATC Codes
Text based similarity	1,661	3,007
Taxonomy based similarity	1,661	3,007
Molecular structure based Similarity	8,738	3,512

Note that from the total number of drugs (10,562), just 2,287 has a **non-empty ATC Code**. In addition, in this experiment we are not using all but just 8,738 drugs, so the number of drugs with non-empty ATC Code are even less, specifically, **2,003 drugs**.

GROUND TRUTH BASED EVALUATION

TEXTUAL SIMILARITY

Number of components for LSA	100	200	500
Pairs in ground truth	97	97	97
Pairs in computed similarity 65	65	65	
Kendall's τ	0.2327	-0.0269	0.0125
Pearson's Correlation	0.7920	0.7385	0.6875
Accuracy	0.7385	0.7385	0.7385
Recall	0.0556	0.0556	0.056

TAXONOMIC SIMILARITY

Graph	Unweighted	Weighted
Pairs in ground truth	97	97
Pairs in computed similarity	65	65
Kendall's τ	0.2212	0.0673
Pearson's Correlation	0.6721	0.6998
Accuracy	0.7538	0.7692
Recall	0.7222	0.7778

MOLECULAR SIMILARITY

Sort of Fingerprint	ECFP	MACCS
Pairs in ground truth	97	97
Pairs in computed similarity	96	96
Kendall's τ	-0.0404	0.0601
Pearson's Correlation	0.8886	0.9186
Accuracy	0.7708	0.8854
Recall	0.12	0.76

The **Clustering evaluation** has provided lights and shadows, while in some cases we have been able to cluster properly the drugs based on their ATC Codes, we have not in several cases. This does not strongly implies our similarity measures are not good. Spectral Clustering, used in this work, and graph-based semi-supervised learning algorithms, in general, are well known to be sensitive to how graphs are constructed from data. In particular if the data has proximal and **unbalanced** clusters these algorithms can lead to poor performance.

On the other hand, some **promising results** have been found in the evaluation based on the ground truth, specially, for the similarity based on Molecular Structure. Nevertheless, the results are not definitive, a need of a **larger ground truth** is clear.