D3.3

Integration of INtERAcT, MelanomaMine and LimTox and application to biomedical publications on paediatric cancers

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D3.3/ 1.0
Work package contributing to the deliverable	WP3
Due date	November, 2021 – M35
Actual submission date	30th November, 2021

Responsible organisation	IBM
Editor	Matteo Manica
Dissemination level	PU
Revision	1.0

Abstract	We describe the development of INtERAcT, its application to paediatric cancer literature, and its integration with implementation previously reported in D3.2.
Keywords	Natural Language Processing, word embeddings, text mining, network medicine

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Matteo Manica (IBM)

Contributors (ordered according to beneficiary numbers)

Davide Cirillo (BSC)

Disclaimer

Executive Summary

This document reports on the integration of INtERAcT in the text mining workflow implemented in the context of D3.2. The workflow has been built to adapt LimTox and MelanomaMine to paediatric tumour abstracts from PubMed and relies on INtERAcT in its downstream component of inferring molecular associations between entities extracted from unstructured text.

In Chapter 1 we provide a summary of D3.2 and we describe the fundamental concepts of INtERAcT as well as the reasons behind the integration. In Chapter 2 we describe INtERAcT more in details, explaining its application to paediatric tumours in the context of the text mining workflow implemented, describing the integration with the iPC text mining workflow. In Chapter 3 we comment on the results obtained in D3.2, describing how to reproduce them and reporting statistics for the networks estimated. In Chapter 4 we report summary and conclusions.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Short summary of D3.2

D3.2 describes a text mining workflow that we implemented for distilling biomedical concepts and entities from PubMed abstracts on paediatric tumours. We considered as starting points two text mining tools developed in BSC: LimTox (Cañada A, 2017) (http://limtox.bioinfo.cnio.es/) and MelanomaMine (http://melanomamine.bioinfo.cnio.es/, https://github.com/cirillodavide/melanomamine). Both tools have been adapted to parse PubMed abstracts and assemble a paediatric tumour-specific corpora as well as extract entities and annotations for downstream NLP and RE tasks. The processed entities have been then used to compute context-specific networks using INtERAcT leveraging the word embeddings learned on the relevant literature identified via LimTox and MelanomaMine.

1.2 Mining biomedical data with INtERAcT

INtERAcT has been conceived as a mean to better leverage learned word vector representations. The major advantage given by its application lies in its unsupervised nature. By exploiting the metric it defines we can easily build relational networks by scoring arbitrary pairs of entities. The only requirements are to initially cluster the vector space learned on the corpus of interest and provide an entities list to be considered (more details in Chapter 2). This allows for a seamless integration with the tools developed in BSC (LimTox and MelanomaMine) at two levels: first, considering word vectors specifically trained on paediatric cancer literature; second, building the relational networks on the context-specific entities extracted with the workflow. To ease its consumption INtERAcT has been made available via a web interface (https://ibm.biz/interact-aas) and as a pip-installable python package (https://github.com/drugilsberg/interact).

1.3 Integration motivation

As shown in the journal paper presenting INtERAcT (Matteo Manica, 2019), standard metrics fail in high dimensions and are not able to successfully reconstruct relevance networks of bio-entities from vector representation of words, especially when working on small corpora. In the context of iPC, it was critical to consider methods that could build robust networks with limited data availability. Hence the idea of integrating INtERAcT in the NLP workflow to compute similarity scores (see Figure 1) and build robust networks of bio-entities for five paediatric tumours of interest.

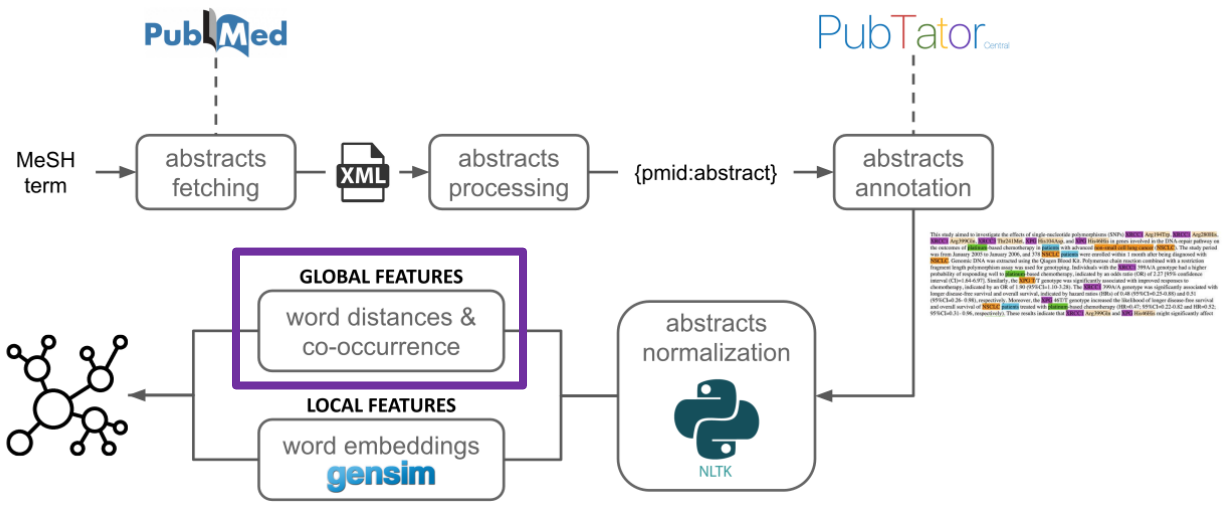

Figure 1: iPC text mining workflow where the role of the INtERAcT integration is highlighted with a purple box.

Chapter 2 INtERAcT

2.1 INtERAcT description

Recently, freely available biomedical corpora are growing enormously, resulting in a remarkable source of knowledge in the form of unstructured text. The large volume of published literature renders impossible an exhaustive manual curation and annotation. Hence the demand for performant methodologies that can analyze text resources, organize them in structured knowledge and learn meaningful representations with no or limited supervision.

INtERAcT (Matteo Manica, 2019) infers relations between concepts, e.g., molecular entities, extracted from the text relying on a completely unsupervised procedure that leverages word embeddings learned via automatic text mining. INtERAcT defines a metric that acts on the vector space of word representations to estimate an interaction score between two entities, e.g., genes or proteins.

In a nutshell (see

Figure 2), by partitioning with a clustering algorithm the word vector space (e.g., K-means), it is possible to define a distance based on the distribution of the cluster assignments of the neighbors of a given word. Considering two words of interest we can define a score based on the Jensen-Shannon divergence (Endres & Schindelin, 2003) of the two neighbors' distributions (detailed derivation and mathematical formulation in the manuscript available here: https://rdcu.be/cAncb).

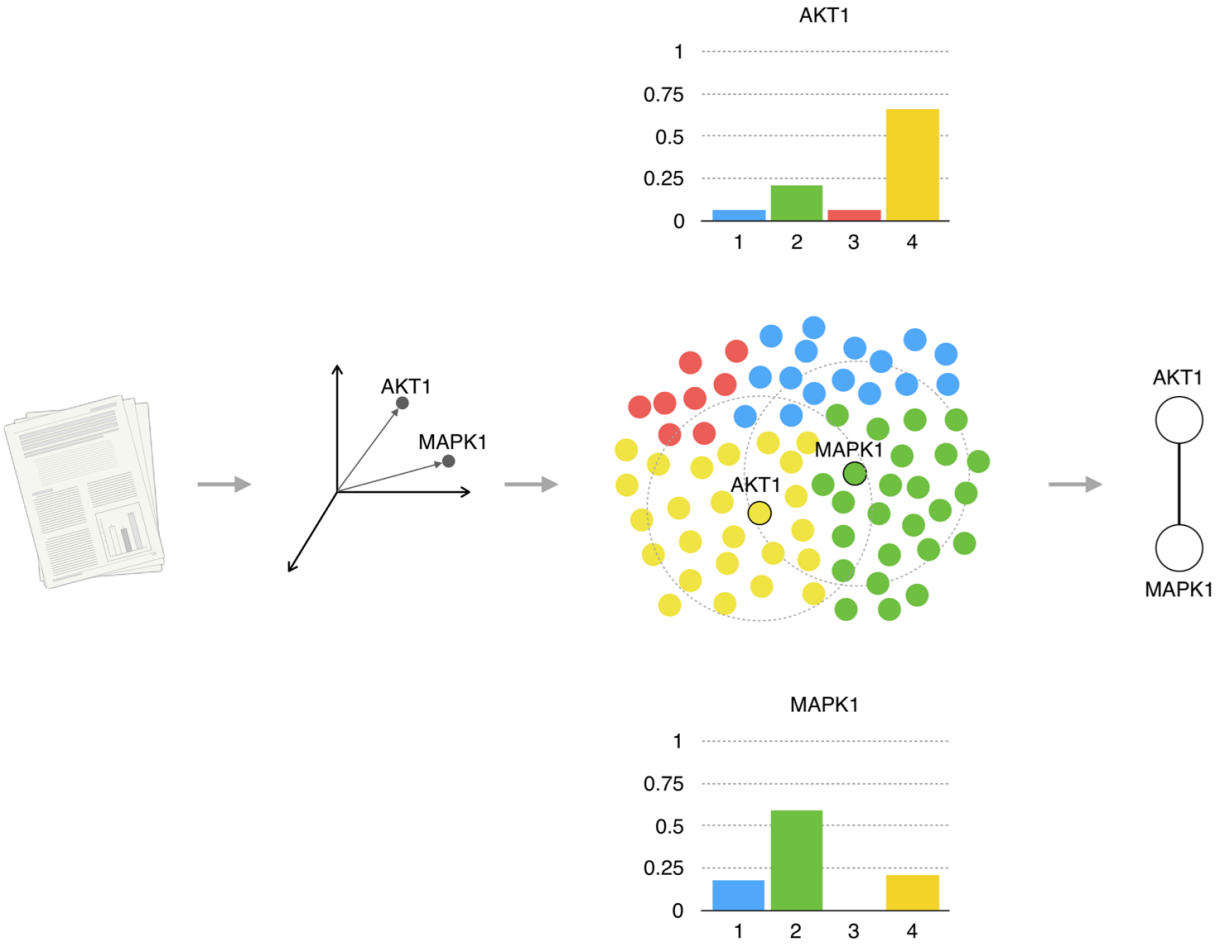

Figure 2: Depiction of INtERAcT. Vector representations of words are used to define neighbours' distributions that are then compared for similarity using a score based on the Jensen-Shannon divergence.

For efficiency, a *k*-d tree (Bentley, 1975) is used to allow for fast retrieval of neighbors in the word embedding space. This guarantees optimal performance when screening candidate word pairs representing bio-entities of interest.

This recipe offers the possibility to quickly screen pairs and, as demonstrated in the publication, to overcome issues that arise from considering standard metrics in high dimensional spaces.

2.2 Integration scheme and INtERAcT application to paediatric cancer literature

Given the flexibility and generality of the methodology, no specific adaptation of INtERAcT was required for the paediatric tumour use-case.

It has been trivial to build paediatric cancer-specific networks starting from the pipeline developed for D3.2. Considering word vectors trained on the corpora identified using the adapted versions of LimTox and MelanomaMine, it was possible to easily define networks of genes, considering a full list of gene names parsed from UniProt (Consortium, 2021).

The full implementation of the pipeline is available here and can be reproduced solely relying on the files present in the repository.

Chapter 3 Integration of INtERAcT and previous implementations

3.1 Results

Following the analysis performed in D3.2, we reconstructed networks focusing on the tumour types reported in Table 1.

Table 1: Tumour types considered reporting the MeSH terms used and the number of abstracts used in the NLP workflows (abstract from February 2020, from D3.2).

Paediatric tumor	MeSH term	Number of abstracts
Neuroblastoma	D009447	23,122
Acute Lymphoblastic Leukaemia	D054198	22,898
Medulloblastoma	D008527	5,148
Ewing's sarcoma	D012512	5,024
Hepatoblastoma	D018197	1,557

For all the five tumour types we run INtERAcT using the optimal configuration tested in the original publication (Matteo Manica, 2019), namely for the partition/clustering we considered 500 clusters and for the neighbours distributions we used the 2000 closest points to the query word (gene name of interest) in the word embedding space.

The relevance networks estimated are obtained after intersecting all gene names retrieved from UniProt with the vocabulary of the corpus of interest. After this filtering steps, INtERAcT compute the pairwise interaction scores for all the filtered entities relying on the *k*-d tree constructed on the full word embedding space.

In Table 2, we report the statistics for the resulting networks estimated with INtERAcT.

Table 2: Statistics for the networks estimated for the five tumour corpora considered in D3.2.

Paediatric tumor	Number of edges	Number of vertices
Neuroblastoma	106,030	459
Acute Lymphoblastic Leukaemia	76,245	389
Medulloblastoma	19,306	195
Ewing's sarcoma	13,695	164
Hepatoblastoma	5,460	103

The generated networks are available in the INtERAcT repository under the iPC example section: https://github.com/drugilsberg/interact/tree/master/examples/ipc. Complementary files, such as the word embedding model trained on the corpus and the embedding matrix, are also available for each tumour type.

Chapter 4 Summary and Conclusion

Herein, we presented the application of INtERAcT to five paediatric tumor types as well as the integration with LimTox and MelanomaMine into the iPC text mining pipeline that has been built in the context of D3.2.

We described the foundational concepts behind INtERAcT and described how it is intrinsically suitable for dealing with contexts where there is limited availability of literature resources, e.g., paediatric tumours. We also presented in which pipeline stage INtERAcT has been applied and how to reproduce the results presented in D3.2 that have been complemented with the generated. Interaction networks statistics in the present deliverable.

The networks produced with INtERAcT, that are the main outcome of the iPC text mining pipeline, will play a key role for the network-based model developed in the downstream work packages. Especially, in the light of the fact that they will integrate the networks estimated at the omic level, bringing in knowledge distilled from all the relevant literature retrieved considering matching MeSH terms.

List of Abbreviations

Abbreviation	Translation
NLP	Natural Language Processing
RE	Relation Extraction
INtERAcT	Interaction networks from vector representation of words

Bibliography

Matteo Manica, R. M. (2019). Context-specific interaction networks from vector representation of words. *Nature Machine Intelligence*, 181–190.

Cañada A, C.-G. S. (2017). LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Research*, W484–9.

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 1858–1860.

Consortium, T. U. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D1.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 509–517.