RZ 3918 (# ZUR1801-015) 01/09/2018
Computer Science/Mathematics 19 pages

Research Report

INtERAcT: Interaction Network Inference from Vector Representations of Words

Matteo Manica, Roland Mathis, Maria Rodriguez Martinez

IBM Research – Zurich
8803 Rüschlikon
Switzerland

Research
Africa • Almaden • Austin • Australia • Brazil • China • Haifa • India • Ireland • Tokyo • Watson • Zurich

INtERAcT: Interaction Network Inference from Vector Representations of Words

Matteo Manica[1,2,*], Roland Mathis[1,*], María Rodríguez Martínez[1, †]

{tte,lth,mrm}@zurich.ibm.com

[1] IBM Research Zürich,
Säumerstrasse 4, 8803 Rüschlikon, Switzerland

[2] ETH Zürich,
Institute für Molekulare Systembiologie, Auguste-Piccard-Hof 1
8093, Zürich, Switzerland

[*] Shared first authorship

[†] Corresponding author

Abstract

Background

In recent years, the number of biomedical publications made freely available through literature archives is steadfastly growing, resulting in a rich source of untapped new knowledge. Most biomedical facts are however buried in the form of unstructured text, and their exploitation requires expert-knowledge and time-consuming manual curation of published articles. Hence the development of novel methodologies that can automatically analyze textual sources, extract facts and knowledge, and produce summarized representations that capture the most relevant information is a timely and pressing need.

Results

We present INtERAcT, a novel approach to infer interactions between molecular entities extracted from literature using an unsupervised procedure that takes advantage of recent developments in automatic text mining and text analysis. INtERAcT implements a new metric acting on the vector space of word representations to estimate an interaction score between two molecules. We show that the proposed metric outperforms other metrics at the task of identifying known molecular interactions reported in publicly available databases for different cancer types.

Conclusions

Our findings suggest that INtERAcT may increase our capability to summarize the understanding of a specific disease or biological process using published literature in an automated and unsupervised fashion. Furthermore, our approach does not require text annotation, manual curation or the definition of semantic rules based on expert knowledge, and hence it can be readily and efficiently applied to different scientific domains, enabling the automatic reconstruction of domain-specific networks of molecular interactions.

Key words: Natural Language Processing, word embeddings, protein–protein interactions, knowledge extraction, prostate cancer.

1 Background

As the number of scientific publications continues to grow exponentially, search engines such as PubMed[1] provide an unprecedented amount of information in the form of unstructured written language. With the accelerating growth of available knowledge – particularly in the biomedical literature – and the breakdown of disciplinary boundaries, it becomes unfeasible to manually track all new relevant discoveries, even on specialized topics. As an example, recent advances in high throughput experimental technologies have yielded extensive new knowledge about molecular interactions in the cell; however most of this knowledge is still buried in the form of unstructured textual information only available as written articles.

As of October 2017, PubMed comprises more than 27.8 million references[2] consisting of biomedical literature from MEDLINE, life science journals, and online books. Most references include links to full–text content from PubMed Central® (PMC[3]) – a free full–text archive of biomedical and life sciences journal literature – or publisher web sites. Currently 14.2 million PubMed articles have links to full–text articles, 4.2 million of which are freely available. The numbers remain high even when focusing on specific fields such as prostate–cancer. For instance, a simple query[4] for prostate–cancer related papers on PMC returns 143321 publications[5]. While a fraction of the information currently available in biomedical publications can be extracted from public databases, the rate at which new research articles are published greatly exceeds the rate at which this information can be currently processed, resulting in an ever wider gap between available knowledge and easily accessible information, e.g. information stored in a database. Clearly the development of novel methodologies that can automatically analyze textual sources, extract facts and knowledge, and produce summarized representations that capture the most relevant information are needed more than ever.

We present here a novel approach to automatically extract knowledge from biomedical publications. Specifically, we focus on the problem of identifying and extracting Protein–Protein Interactions (PPIs) from a disease–specific text corpus and building an interaction network. While our approach is generic and can be applied to any knowledge domain, we demonstrate its strength using the biomedical literature related to prostate–cancer (PC), a complex disease with multi–factorial etiology. PC is the second most common cancer type and the fourth leading cause of cancer death in men worldwide Ferlay et al. [2015]. Despite the large number of newly diagnosed cases, the majority of cases in older men are clinically insignificant, meaning that the life expectancy of the patient is shorter than the time required by the disease to manifest any symptoms Zlotta et al. [2013]. However a small fraction of new cases are aggressive cancers that require intervention. The current prognostic factors are not sufficient to precisely stratify these two groups Cooperberg et al. [2009], and thus PC is prone to overdiagnosis and treatment associated with debilitating side effectsR et al. [2011].

While various approaches to automatically extract PPIs information from unstructured text are already available, many of these methods require feature engineering and expert-domain knowledge for good performance, hence preventing full automation. Commonly proposed methodologies exploit machine learning approaches Tikk et al. [2010], Tjioe et al. [2010], data mining tools Mandloi and Chakrabarti [2015], co-occurrences Barbosa-Silva et al. [2011], Fleuren et al. [2013], Raja et al. [2013], Usie et al. [2014], or rules–based text mining Torii et al. [2015].

Recently, word embedding techniques based on deep learning have been proposed as a more advanced approach to process textual information in an unsupervised fashion. Word embedding is a term used to identify a set of methods for language modelling and feature learning, where words in a vocabulary are mapped into vectors in a continuous, high dimensional space, typically of several hundred dimensions Collobert and Weston [2008]. In this representation, words that share a similar context in the corpus are located in close proximity in the word embedding vector space. Besides representing words' distributional characteristics, word–vectors can capture the semantic and sequential information of a word in a text, providing a richer vector representation

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`

[2] The current size of the database can be obtained by typing "1800:2100[dp]" into the search bar.

[3] `https://www.ncbi.nlm.nih.gov/pmc/`

[4] `https://www.ncbi.nlm.nih.gov/pmc/?term="prostate+cancer"`

[5] Number obtained as of 12 October 2017

than frequency–based approaches. Word–vector representations have gained broad recognition thanks to the recent work of Mikolov et al. Mikolov et al. [2013a,b], who demonstrated that word embeddings can facilitate very efficient estimations of continuous–space word representations from huge datasets ($\sim$ 1.6 billion words).

Since this seminal work, word embeddings based on neural networks have been used to address different tasks of natural language processing. For instance, word embeddings were used in Nie et al. [2015] for the task of event trigger identification, i.e. to automatically detect words or phrases that typically signify the occurrence of an event. In Zhou et al. [2014], a combination of features extracted from a word embedding plus syntactic and semantic context features was used to train a support vector machine classifier for the task of identifying event triggers. Such approaches have been shown to be efficient in identifying the semantic and syntactic information of a word and incorporate it into a predictive model. Word embeddings have also been used as token features – semantic units of words and characters extracted from a corpus for further processing – to extract complete events represented by their trigger words and associated arguments Li et al. [2015]; to build knowledge regularized word representation models that incorporate prior knowledge into distributed word representations for semantic relatedness ranking tasks Wang et al. [2015]; and to simultaneously analyze the semantic and contextual relationship between words Jiang et al. [2016]. Finally, alternative deep learning approaches based on autoencoders and a deep multilayer neural network have been used to extract PPIs, where the features were extracted by a Named Entity Recognition module coupled to a parser and principal component analysis Zhao et al. [2016].

While these approaches have shown the versatility of word embeddings to support text analysis through current natural language processing (NLP) tools, approaches that can automatically extract molecular interactions from unstructured text in a completely unsupervised manner are still missing. To bridge this gap we present our methodology hereby referred as INtERAcT (Interaction Network infErence from vectoR representATions of words). Our approach can be summarized as follows. We first create a word embedding from a corpus of freely available publications related to prostate–cancer. Next, we cluster the learned word–vectors in the embedded word–space and find groups of words that convey a close semantic and contextual similarity. Then we focus on proteins and predict PPIs using a novel similarity measure based on the Jensen–Shannon divergence. To demonstrate the generalization potential of our approach to other domains of knowledge, we repeat the exercise and apply INtERAcT to a corpus of publications related to 10 different cancer types, and validate our results using STRING[6] database Szklarczyk et al. [2015] as a benchmark.

2 Results

2.1 Applying INtERAcT on prostate–cancer publications

Building a word embedding specific to prostate–cancer: In the following section we describe the application of INtERAcT to the problem of reconstructing a prostate–cancer pathway. Text pre-processing and building of the word embedding follows the methodology described in Section 5. Briefly, a text corpus is assembled by downloading the XML version of $\sim$ 140000 PubMed Central publications matching the query *"prostate cancer"*. Only abstracts are processed, as they provide a concise and clean summary of the article's main findings. Rare words and bi–grams occurring less than 50 times in the corpus are removed. The remaining sentences are tokenized – segmented into linguistic units – and used to build a word embedding. After processing (see section 5.1), our dictionary is composed of $\sim$ 21000 single words and common bi–gram, e.g. prostate_cancer, cell_proliferation and gene_expression. Using this dictionary, we build a word embedding using a vector representation of 500 dimensions and a context window of size 8 words (4 words to the right and the left of each target word). See Section 5 and Fig. 4 for details.

Applying INtERAcT: The results of the embedding are clustered into groups conveying similar semantic meaning using K–means with 500 clusters. We next identify the k-nearest neighbors of each protein as described in section 5. The neighborhood size is set to k =2000 in order to keep a balanced trade-off between number of neighbors and number of clusters. We use the cluster assignment distribution of selected words to calculate the pairwise similarity scores based on the

[6] `https://string-db.org/`

JSD as shown in 5. This last step is done on a subset of words – in this example, a list of molecular entities defined using UniProt. We interpret this JSD-based distance metric as the likelihood of a PPI. See Section 5 and Fig. 1 for details.

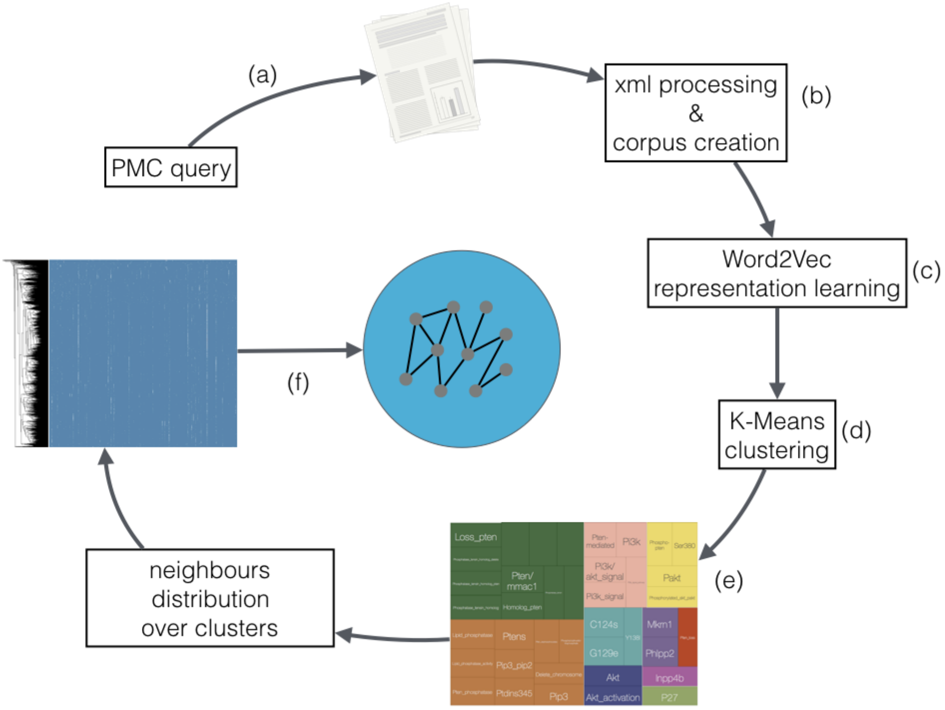


Figure 1: **Schematic representation of INtERAcT**. **(a)** PubMed Central was searched using *"prostate cancer"* as a keyword. Close to 140000 publications were retrieved and downloaded in XML format. **(b)** The XML files were processed to remove non informative words and characters such as stop–words, punctuation and isolated numbers. **(c)** The resulting processed sentences constituted a corpus that was used to train a word embedding vector space of 500 dimensions. **(d)** All the words in the embedding where clustered using K–means with 500 clusters. **(e)** A visualization of the word neighbourhood of PTEN – a gene often mutated in prostate–cancer. For visualization purposes, related words are grouped and colored according to the clustering. **(f)** A list of words – e.g. protein comprised in a prostate–cancer pathway – were selected from the word embedding, and its pairwise distance calculated by comparing neighborhood distributions using the Jensen–Shannon divergence. An interaction graph was built using the similarity scores.

To benchmark the inferred network we focus on the list of molecular entities reported in the prostate–cancer pathway as defined by the Kyoto Encyclopedia of Genes and Genomes[7] (KEGG) and apply INtERAcT to the task of reconstructing the connectivity between these entities. Out of the 87 molecular entities that constitute the KEGG pathway, 67 are found in the embedding, and thus can be used as a validation set.

We interrogate INtERAcT and query the interactions between the 67 proteins of our validation set. Fig. 2a graphically shows the top–50 inferred interactions in our prostate–cancer gene validation set. The full set of interactions with similarity scores can also be found as a table in the Supplementary Material S1. Please, notice that while KEGG provides a well-established reference for function-specific pathways, KEGG merges gene family members in a single node-entity (e.g. AKT1, AKT2 and AKT3 become AKT), and hence a direct comparison between KEGG prostate–cancer pathway and INtERAcT inferred results is not possible.

[7] `http://www.genome.jp/dbget-bin/www_bget?pathway+map05215`

Comparing INtERAcT to STRING: In order to assess the quality of our predictions, we use STRING[8] database Szklarczyk et al. [2015] as a benchmark. STRING is a comprehensive protein interaction database currently including experimental data from DIP[9]Salwinski et al. [2004], BioGRID[10]Chatr-aryamontri et al. [2017], IntAct[11]Hermjakob et al. [2004], and MINT[12]Licata et al. [2011], and curated data from BioCyc[13]Caspi et al. [2007], GO [14]Ashburner et al. [2000], KEGG[15]Kanehisa and Goto [2000]Kanehisa et al. [2016], and Reactome[16]Croft et al. [2014]. STRING provides a confidence score that integrates information about genomic proximity, gene fusion events, phylogenetic co–occurrences, homology, co–expression, experimental evidence of interaction, simultaneous annotation in databases and automatic text–mining Franceschini et al. [2013]. Importantly for the sake of comparing STRING and INtERAcT results, STRING text–mining is done by using a combination of co-occurrences and natural language processing based on a rule-based system Šarić et al. [2006].

To quantitatively evaluate the goodness of INtERAcT predictions, we employ the Area Under a receiver operating characteristic Curve (AUC metric Florkowski [2008]) using STRING interactions as a ground truth, and compare our JSD-based score (Eq. 5) with other similarity scores commonly used in the literature, namely scores based on cosine and Euclidean distance. Figure 2b reports a summary of our findings. The ROC curve for the INtERAcT score (orange curve), a cosine–based distance score (blue line) and an Euclidean–based distance are comparatively shown. INtERAcT achieves a 0.70 AUC, significantly better than the cosine distance, which achieves a 0.61 AUC. The Euclidean based distance measure performs practically equivalent to a random predictor with an AUC value of 0.50. This poor performance is expected as the Euclidean distance, and more generically, L_k norms, tend to map pairs of points to uniform distances in high dimensional spaces Aggarwal et al. [2001]. The curves' trends reinforce the intuition that a neighborhood–aware metric is better able to capture functional associations from unstructured text than methods that limit their analysis to the positions of word–vectors in the embedding.

As an additional measure of agreement, we also compute the rank correlation between INtERAcT and STRING scores. To compute the correlation values, all predicted interactions by INtERAcT and STRING were used without applying any confidence cut-off. The INtERAcT and STRING scores (as downloaded on 19/10/17) used to compute the correlations are provided as additional supplementary tables. The resulting correlation value is positive and very significant ($\rho = 0.31$, $p = 1.6e^{-42}$), and is higher compared to the correlation obtained using cosine and Euclidean distance–based metrics ($\rho = 0.29$, $p = 4.1e^{-39}$ and $\rho = 0.19$, $p = 1.25e^{-16}$ respectively). INtERAcT outperforms again the cosine and Euclidian distance–based metrics. We note that while the correlation values obtained for INtERAcT and the cosine-based scores seem to be relatively close, their difference turns out to be highly significant with a p-value of $p = 1.99e^{-08}$ when the number of interaction scores used to compute the correlations is taken into account (number of interactions = 132357). The significance of the difference of two correlation values can be computed using the Fisher z-transformation Fisher [1915], which transforms the Spearman correlation values into normally distributed variables whose difference can be evaluted used a standard t-test.

2.2 Applying INtERAcT on other cancer pathways

We next focus on investigating the generalization of INtERAcT to other knowledge domains. For this task, we extend our analysis to nine additional cancer types: acute myeloid leukemia, bladder cancer, chronic myeloid leukemia, colorectal cancer, glioma, small cell lung cancer, non–small cell lung cancer, pancreatic cancer and renal cell carcinoma. The gene sets for each cancer type are taken from their respective cancer–specific pathway as annotated in KEGG. These cancer types are selected according to two criteria: first, there is a cancer-specific KEGG pathway to define

[8] https://string-db.org/

[9] http://dip.doe-mbi.ucla.edu/dip/

[10] https://thebiogrid.org/

[11] http://www.ebi.ac.uk/intact/

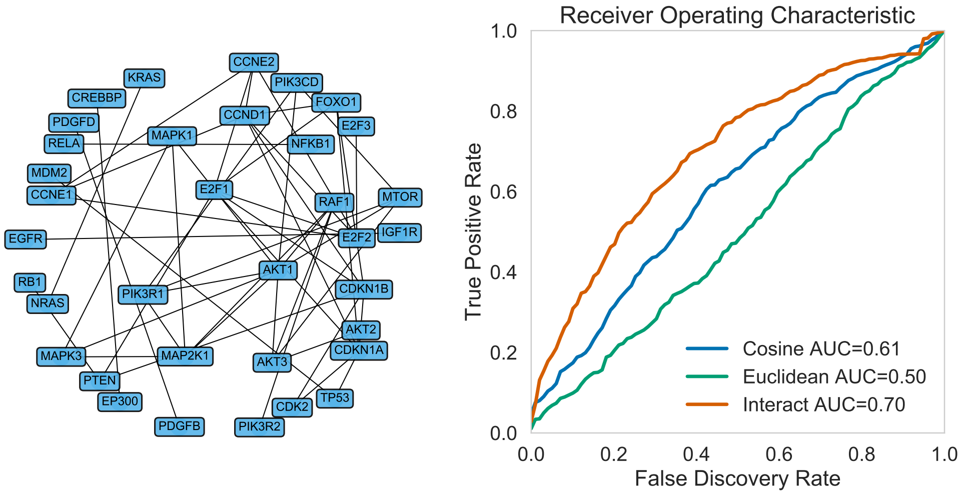
[12] http://mint.bio.uniroma2.it/

[13] https://biocyc.org/

[14] http://www.geneontology.org/

[15] http://www.kegg.jp/

[16] http://www.reactome.org/



(a) (b)

Figure 2: **(a) Top 50 prostate–cancer protein–protein interactions inferred by INtERAcT**. The prostate–cancer gene set was defined according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) prostate–cancer pathway, and includes molecular entities known to be important in prostate–cancer onset and development. The interactions and associated scores were computed using a word embedding trained on ∼140000 prostate–cancer abstracts freely available on PubMed Central and INtERAcT, our proposed methodology to extract functional interactions from a word embedding. **(b) Performance of INtERAcT on a prostate–cancer gene validation set compared to other distance measures using STRING as a ground truth.** We used a ROC (Receiver Operating Characteristic) curve to quantify the accuracy of the inferred interactions in a set of prostate–cancer-related genes. INtERAcT (orange curve) significantly outperforms alternative, commonly used metrics on a word embedding such as a cosine distance–based similarity (blue curve) and a similarity score based on the Euclidean distance (green curve).

a gene set, and second, we could retrieve at least 10000 cancer-specific publications in PubMed Central. The second criterion is needed in order to obtain a corpus size that guarantees a good reconstruction of the word–vectors when building the word embedding. We then defined new query words specific to each cancer type and repeated the procedure described in 2.1. The full list of used query words for each cancer type can be found in the Supplementary Material (section 1.7),

In Figure 3 we report the median ROCs for three different distance metrics: INtERAcT (orange curve), cosine (blue curve) and Euclidean (green curve) metrics. In order to obtain a confidence for the curves using the different pathways considered, we built empirical confidence intervals (CIs). The CIs at level 68% are reported (one standard deviation from the mean) in Figure 3. The CIs are generated performing an empirical bootstrap on the different pathways. For each FPR (False Positive Rate) level 5000 values are sampled with replacement from the TPR (True Positive Rate) values obtained from the different pathways to generate an empirical distribution and build the intervals.

Finally, we compare the similarity of scores predicted by INtERAcT and STRING by computing the Spearman rank correlation between both sets of scores. The values are shown in Table 1. For all analyzed pathways, the correlation is positive with a strongly significant p-value (the p-values range from 10^{-06} to 10^{-48}). The correlation between the negative logarithm of the p-value and the number of publications is 0.87, revealing that the main factor determining the significance of the p-value is the number of publications used to build the embedding.

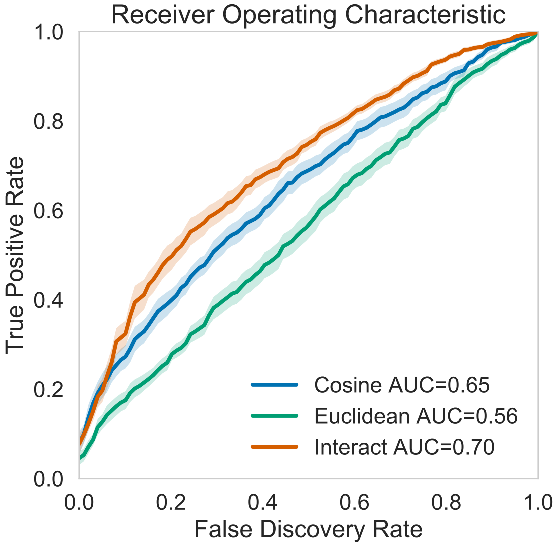


Figure 3: **INtERAcT performance compared to other distance measures using STRING as a ground truth.** We use ROC (Receiver Operating Characteristic) curves to quantify the quality and performance of inferred interactions. The curves here reported refer to the inference performed on the KEGG cancer pathways considered in the analysis. Using naive approaches such as a similarity based on the Euclidean distance (green curve) between word–vectors led to poor results. Other methods such as cosine–based similarity (blue curve) showed an improvement. INtERAcT (orange curve) achieved the best performance predicting interactions reported in STRING. The confidence intervals (CIs) at level 68% are reported (one standard deviation from the mean). To generate the empirical distribution we used sampling with replacement at different false positive rates (FPRs) of the true positive rates (TPRs) given by the different pathways. The confidence intervals reported are at level 68% (one standard deviation from the mean)

3 Discussions

Our findings suggest that while having a large enough corpus is of paramount importance to obtain robust predictions, the number of publications seems to play a moderate role in determining the strength of the association between INtERAcT and STRING scores (see Table 1). For instance, the highest correlation value 0.47 is found in colorectal cancer, which has the second highest number of publications used to build the embedding. However, prostate–cancer only shows a moderate correlation of 0.31, while having the largest number of publications used. We hypothesize that while having a large corpus of publications is beneficial to build a high–quality embedding, very active fields of research where a high number of publications are available may also be prone to having a high rate of *noisy publications*. Here noise can take the form of low-quality publications that report inconsistent results, or studies based on high-throughput analyses with a high false discovery rate. We also note that in taking STRING as ground truth we are implicitly absorbing its false and true discovery rates into our error rates. For instance, interactions reported by STRING that might occur in a different context but not in cancer (e.g. mouse interactions not occurring in cancer) will get penalized as false negatives if INtERAcT correctly predicts them as a non–interaction.

Taken all together and within the limitation of not having an unbiased ground truth to evaluate our predictions, INtERAcT shows a good agreement with the information reported by STRING. Our results indicate that our unsupervised approach is able to recapitulate to a large extent the knowledge obtained through manual curation of scientific literature.

	Correlation	Number of Proteins	Papers	p-value
KEGG Acute Myeloid Leukemia	0.340172	34	34532	2.993059e-14
KEGG Bladder Cancer	0.359215	30	35331	2.923255e-13
KEGG Chronic Myeloid Leukemia	0.337301	23	14247	1.799958e-07
KEGG Colorectal Cancer	0.465745	48	118336	2.559023e-48
KEGG Glioma	0.256279	36	64712	6.565343e-10
KEGG Small Cell Lung Cancer	0.267501	28	32233	2.064622e-06
KEGG Non Small Cell Lung Cancer	0.279969	31	67048	9.132340e-09
KEGG Pancreatic Cancer	0.349509	47	62668	2.276079e-26
KEGG Prostate Cancer	0.312031	67	132357	1.679381e-42
KEGG Renal Cell Carcinoma	0.427098	30	37169	1.925472e-15

Table 1: **INtERAcT–STRING rank–correlation on KEGG's cancer pathways.** The table reports the Spearman correlation values of INtERAcT predictions and STRING confidence scores. For all analyzed pathways and cancer types, INtERAcT– and STRING–derived scores show a positive and significant correlation.

4 Conclusions

We have presented a fully unsupervised method to automatically extract context–specific molecular interaction networks from freely available publications, without any doubt, the fastest growing source of scientific information. Our approach does not require time–consuming manual curation nor labelling of the text. Indeed, no annotations or other manual processing step are required. Furthermore, the results presented here have been obtained without optimization of hyper–parameters.

We have described the steps to reconstruct a context–specific pathway from prostate-cancer publications. When comparing the inferred interactions to STRING, our method outperforms other scores built on commonly used metrics (cosine and Euclidean metric). On a more extensive validation on multiple cancer pathways, the results remain consistent and we have a significant agreement on the information reported by STRING. We would like to highlight that STRING predicts interactions using a combined score that integrates information from many disparate data sources including genomic proximity, gene fusion events, phylogenetic co–occurrences, homology, co–expression, experimental evidence of interaction, simultaneous annotation in databases and automatic text–mining. Text–mining is done using a combination of co-occurrences and natural language processing based on a rule-based system Šarić et al. [2006]. Our methodology on the other hand is a completely unsupervised approach only based on publications that does not require either expert–knowledge or rules setting. When focusing on reconstructing a prostate–cancer pathway, we achieved a 0.70 AUC score using STRING as benchmark. We notice that the choice of benchmark is likely overpenalising the evaluation of the precision and recall of our method, as STRING reports many interactions that are not cancer-specific.

We expect the proposed algorithm to be highly relevant for a variety of state of the art text–mining methods. Especially, we are convinced that the proposed methodology can be used to generate hypotheses for detection of biological processes relevant to common and complex diseases and can establish a novel, unsupervised and high–throughput approach to drive drug discovery and advance the frontier of targeted therapies.

5 Methods

In this section we present the elements that constitute INtERAcT and describe the approach adopted to automatically build a network of molecular interactions starting from a domain-specific text corpus.

5.1 Text processing

We begin by using a basic and lightweight pipeline for text processing. First, we filter out non–informative words such as extremely common words (e.g. a, the, and other stop–words), rare words (low occurrence in the corpus), non–informative characters like punctuation or isolated numbers and convert text to lower–case. Please, notice that we only remove isolated numbers in order to leave intact and be able differentiate gene names (e.g. AKT1, AKT2 and AKT3). We next identify bi–grams – sequences of 2 words that often appear together and thus are considered a single entity, e.g. New_York – by summing up the occurrences of two words appearing sequentially together in the corpus and setting a threshold of the minimal number of occurrences. The names of a gene, its aliases and corresponding protein are treated as synonyms and mapped to a single name entity using a dictionary obtained from UniProt[17]. Sentences are generated using an English language tokenizer Bird et al. [2009] – a software used to segment a text into linguistic units, in our case, sentences – before punctuation is removed. The result of this process is a corpus of sentences that can be used for further analysis.

5.2 Word embeddings

Word embeddings are the output of a family of methods that produce, starting from raw text, a real vector representation of each word and phrase contained in the original text corpus. In this work we build the embedding using the Word2Vec implementation proposed by Mikolov et al. Mikolov et al. [2013b], a shallow, two–layer neural network based on a skip–gram model. Briefly, the skip–gram model aims to predict the surrounding words, i.e. the context, of a target word given as an input (see Fig. 4). In practice, a word's context is defined by considering a window of size $2n$ to the left and the right of each target word. Each pair target–context word is then fed into the neural network with a single hidden layer of dimension d that is trained to optimize the probability of predicting context words given a target word as input. It has been reported that the quality of the word embedding increases with the dimensionality of the internal layer that produces the vector representation, d, until it reaches a critical point where marginal gain diminishes Mikolov et al. [2013a]. Hence this parameter has to be appropriately chosen according to the size of the vocabulary and text corpus.

The word embedding learning process is naturally optimized to capture the contextual associations between words: If two words tend to appear in similar contexts, they will be mapped into similar word–vectors. In practice, it has been shown that word embeddings outperform methods based on counting co-occurrences of words on a wide range of lexical semantic tasks and across many parameter settings Baroni et al. [2014].

5.3 Extracting interactions from the embedding

Once the embedding is built, our aim is to design a methodology that can predict PPIs based on the distribution of word–vectors in the word embedding. We exploit the idea that molecular entities that interact with each other and are involved in similar biological processes are likely to appear in similar word contexts, and thus will be mapped to neighboring positions in the word–vector space. It is hence possible to predict functional similarities between molecular entities based on their mapping and proximity in the word embedding.

Our task is therefore to find optimal ways of measuring proximity in the word embedding. A first, obvious approach to define proximity between two word–vectors is to use the Euclidean distance and a distance threshold: molecular entities within this threshold can be considered similar and thus predicted to interact. However, the use of the Euclidean metric, and more generically, the use of L_k norms, is problematic as the high dimensionality of the space can make certain regions of the space too sparse. In addition, in high dimensional spaces L_k norms map points to uniform distances from each other, and hence the concepts of proximity, distance or nearest neighbor are not quantitatively meaningful Aggarwal et al. [2001].

[17] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz

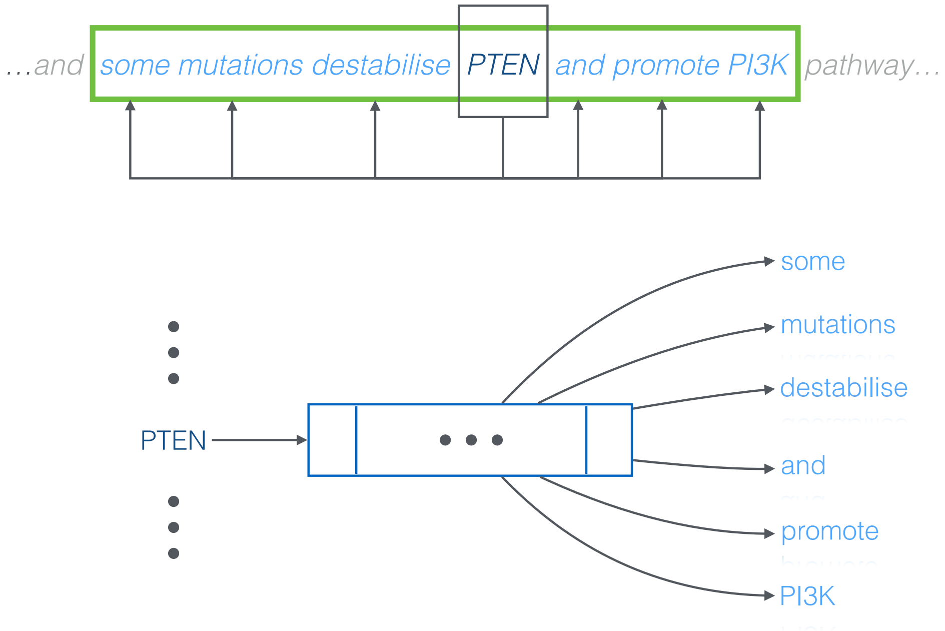


Figure 4: **Description of the skip-gram model** used in Word2Vec to find an optimal representation to predict the surrounding context of a target word. The example highlights the window around PTEN, a gene implicated in many cancer processes. The target word, PTEN, is linked to each of its neighboring words and the pairs are fed into the network. The learning process optimizes the probability of predicting the contextual words of PTEN.

INtERAcT exploits an alternative approach that does not rely on the direct use of L_k norms, but instead defines similarities between words by looking at the semantic meaning of the neighbors. Specifically, we predict PPIs by comparing the neighborhoods of words representing molecular entities. To do so, we first need to cluster the word–vectors of the embedding.
Clustering words: We start by defining $\mathcal{W}$ as the set of n words present in the embedding $\mathcal{E} \in \mathbb{R}^{n \times d}$ where d is the embedding dimension, which corresponds to the dimension of the neural network's hidden layer used to build the embedding. We cluster the word–vectors in the embedding space using a K–means algorithm with C clusters. The number of clusters is chosen according to the vocabulary size in order to have both a fine grained word representation and sufficient number of words per cluster. Each word is hence associated with a cluster according to the following mapping:

$$CL : \mathcal{W} \rightarrow \{1, \dots, C\} \tag{1}$$

.

The obtained clusters group together words that are close in the vector representation space and hence tend to appear in similar contexts in the corpus. These clusters can then be used to build fingerprints of each entity in the embedding and to convey the semantic meaning of a word based on the cluster membership of its neighbors.
Finding nearest neighbors: In order to build word fingerprints, our algorithm requires the identification of the nearest neighbors of each target word. An efficient method to retrieve the topological neighbors without having to compute all pairwise distances at each query is k–d trees, a space–partitioning data structure that can be used to organize points in a k-dimensional space Bent-

ley [1975]. A nearest–neighbor–search can then associate every word in the embedding with a set $\mathcal{N}$ of K nearest neighbors in the embedding:

$$KNN : \mathcal{W} \rightarrow \mathcal{N} \ . \tag{2}$$

The optimal number of neighbors depends on the number of clusters C, and it is chosen as a trade–off between the benefit of having enough cluster assignment variability among the neighbors, while keeping the neighborhood of each word small when compared to the total word count of the embedding. The mapping KNN can be used to efficiently retrieve the shortest paths between two words and identify their nearest words.

Word distribution: We are now able to associate each word in the embedding with a discrete probability distribution that can be computed by analyzing the cluster membership of the nearest neighbors. The number and cluster occupancy of the neighbors can then be interpreted as a discrete probability distribution conveying the semantic meaning of each target word. Furthermore, pair–wise comparisons of these distributions enable us to define similarities between words (see Fig. 5).

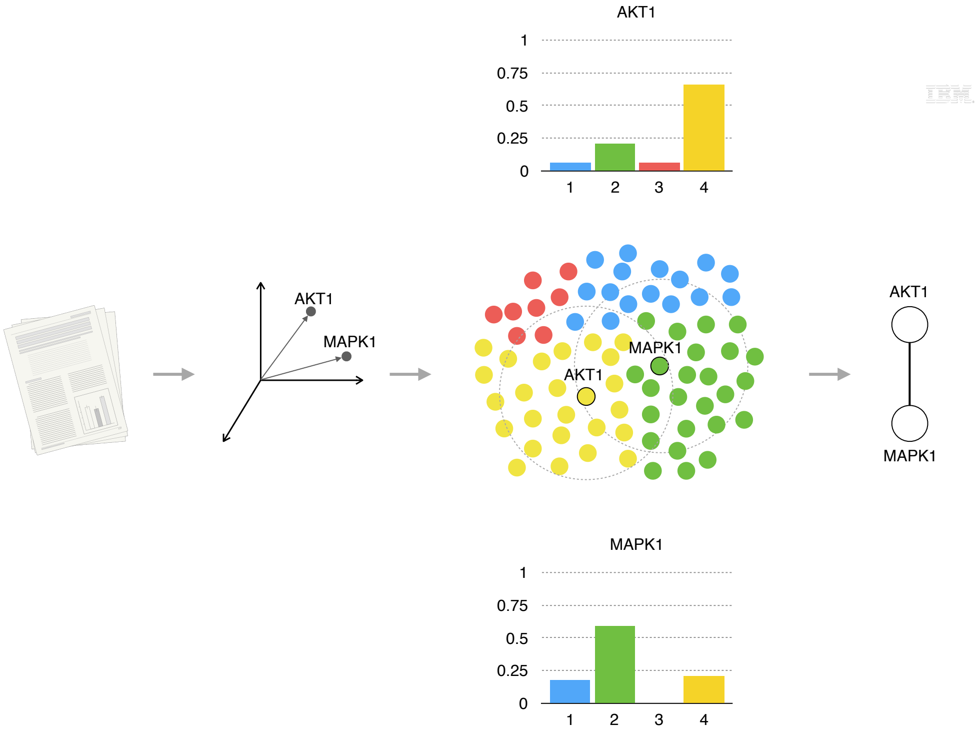


Figure 5: **Schematic representation of INtERAcT.** The parsed text is used as input in a word embedding algorithm. The word–vectors are clustered into groups of similar semantic meaning and the distributions of cluster assignments across clusters are used to compute and predict interactions between molecular entities.

The pseudocode of the described algorithm can be found in the Supplementary Material, section 1. The output of the algorithm is a matrix of probability distributions $\mathcal{D} \in \mathbb{R}^{n \times C}$ where each row contains the cluster assignments of each target word.

Computing similarity scores: We can finally compute the functional association between words of interest by computing the similarities between the neighbors' cluster assignments of protein entities in the embedding. We use a score based on the Jensen–Shannon divergence (JSD), defined as follows:

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3}$$

where $M = \frac{1}{2}(P+Q)$ and D_{KL} is the Kullback–Leibler divergence for discrete probability distributions:

$$D_{KL}(P||Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{4}$$

. The choice of the JSD as a scoring function is motivated by its useful properties. In addition to providing a symmetrized version of the Kullback–Leibler divergence, JSD is a finite value comprised in the interval $[0, \log(2)]$ Lin [1991], the lowest bound being reached when two distributions are identical. Furthermore, the square root of the Jensen–Shannon divergence is a metric Endres and Schindelin [2003], and thus JSD is an appropriate function to capture similarities between distributions.

Here we take advantage of the non–negativity of JSD to define the similarity S_{ij} between words i and j as follows:

$$S_{ij} = \exp(-\alpha JSD_{ij} + \beta) \tag{5}$$

where $JSD_{ij} = JSD(\mathcal{D}_i||\mathcal{D}_j)$ and α and β are a scaling and an offset parameters respectively. In the following, we set the offset parameter $\beta = 0$. Under the transformation defined by Eq. 5, two identical distributions have a score equal to 1 and substantially different distributions (with a divergence close to the JSD upper bound) have a score ~ 0.0. While larger values of α can bring this theoretical minimal value closer to 0, a very high α will make S_{ij} decay too steeply, shrinking the regime where S_{ij} can effectively rank pairs of words according to their similarity. We found that the choice of $\alpha = 7.5$ and $\beta = 0.0$ was empirically efficient at capturing similarities between words given the theoretical bounds for the JSD (see Supplementary Material Figure S1).

Equipped with the similarity score as defined in Eq. 5, we are now in a position to build a weighted interaction graph where the nodes are the chosen entities (proteins in our case) and the edges are weighted by the similarity value of the nodes they connect.

Declarations

Acknowledgments

We would like to acknowledge Dr. Costas Bekas and Dr. Yves Ineichen for useful discussions. The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.

Authors contributions

M.M., R.M. and M.R.M. conceived the study and analyses. M.M. and R.M. implemented INtERAcT and performed data analysis. M.R.M. provided biological analysis and interpretation. M.M., R.M. and M.R.M. wrote the manuscript with input from all authors.

Competing financial interests

The authors declare no competing financial interest.

Availability of data and materials

The article abstracts used to generate INtERAcT protein–protein interaction scores can be freely downloaded from PubMed Central. The article collection as well as access to INtERAcT are also available from the corresponding author on request. For this project, STRING interaction scores were downloaded on 19/10/17. STRING historical data can be downloaded from `https://string-db.org/cgi/access.pl?footer_active_subpage=archive`, or obtained from the corresponding author on request. All other data generated or analysed during this study are included in this published article and its supplementary information files.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, pages 420–434, London, UK, UK, 2001. Springer-Verlag. ISBN 978-3-540-41456-8. URL http://dl.acm.org/citation.cfm?id=645504.656414.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

Adriano Barbosa-Silva, Jean-Fred Fontaine, Elisa R Donnard, Fernanda Stussi, J Miguel Ortega, and Miguel A Andrade-Navarro. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics*, 12:435, November 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-435. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228910/.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1023.

Jon Louis Bentley. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM*, 18(9):509–517, September 1975. ISSN 0001-0782. doi: 10.1145/361002.361007. URL http://doi.acm.org/10.1145/361002.361007.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Ron Caspi, Hartmut Foerster, Carol A Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Seung Y Rhee, Alexander G Shearer, Christophe Tissier, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl_1):D623–D631, 2007.

Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, et al. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, 2017.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. URL http://dl.acm.org/citation.cfm?id=1390177.

Matthew R. Cooperberg, Jeanette M. Broering, and Peter R. Carroll. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *Journal of the National Cancer Institute*, 101(12):878–887, 2009.

David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R. Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–D477, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1102. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965010/.

Dominik Maria Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003. URL http://ieeexplore.ieee.org/abstract/document/1207388/.

Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5):E359–386, 2015. ISSN 1097-0215. doi: 10.1002/ijc.29210.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915. ISSN 00063444. URL `http://www.jstor.org/stable/2331838`.

Wilco WM Fleuren, Erik JM Toonen, Stefan Verhoeven, Raoul Frijters, Tim Hulsen, Ton Rullmann, René van Schaik, Jacob de Vlieg, and Wynand Alkema. Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining. *BioData Mining*, 6:2, February 2013. ISSN 1756-0381. doi: 10.1186/1756-0381-6-2. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3577498/`.

Christopher M Florkowski. Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83–S87, August 2008. ISSN 0159-8090.

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue):D808–D815, January 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1094. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531103/`.

Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue): D452–D455, January 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh052. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308786/`.

Z. Jiang, L. Li, and D. Huang. An Unsupervised Graph Based Continuous Word Representation Method for Biomedical Text Mining. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(4):634–642, July 2016. ISSN 1545-5963. doi: 10.1109/TCBB.2015.2478467.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, January 2000. ISSN 0305-1048. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/`.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1070. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1070`.

Chen Li, Runqing Song, Maria Liakata, Andreas Vlachos, Stephanie Seneff, and Xiangrong Zhang. Using word embedding for bio-event extraction. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). Stroudsburg, PA: Association for Computational Linguistics*, pages 121–126, 2015. URL `http://www.anthology.aclweb.org/W/W15/W15-38.pdf#page=133`.

Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861, 2011.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan 1991. ISSN 0018-9448. doi: 10.1109/18.61115.

Sapan Mandloi and Saikat Chakrabarti. PALM-IST: Pathway Assembly from Literature Mining - an Information Search Tool. *Scientific Reports*, 5, May 2015. ISSN 2045-2322. doi: 10.1038/srep10021. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4437304/`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, January 2013a. URL `http://arxiv.org/abs/1301.3781`. arXiv: 1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA, 2013b. Curran Associates Inc.

Yifan Nie, Wenge Rong, Yiyuan Zhang, Yuanxin Ouyang, and Zhang Xiong. Embedding assisted prediction architecture for event trigger identification. *Journal of Bioinformatics and Computational Biology*, 13(03):1541001, January 2015. ISSN 0219-7200. doi: 10.1142/S0219720015410012. URL `http://www.worldscientific.com/doi/abs/10.1142/S0219720015410012`.

Chou R, Croswell JM, Dana T, and et al. Screening for prostate cancer: A review of the evidence for the u.s. preventive services task force. *Annals of Internal Medicine*, 155(11):762–771, 2011.

Kalpana Raja, Suresh Subramani, and Jeyakumar Natarajan. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database: The Journal of Biological Databases and Curation*, 2013, January 2013. ISSN 1758-0463. doi: 10.1093/database/bas052. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3548331/`.

Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32 (suppl_1):D449–D451, 2004.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue): D447–D452, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1003. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383874/`.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. *PLoS Computational Biology*, 6(7), July 2010. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1000837. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2895635/`.

Elina Tjioe, Michael W Berry, and Ramin Homayouni. Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). *BMC Bioinformatics*, 11(Suppl 6):S14, October 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-S6-S14. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3026361/`.

Manabu Torii, Cecilia N. Arighi, Gang Li, Qinghua Wang, Cathy H. Wu, and K. Vijay-Shanker. RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 12(1):17–29, 2015. ISSN 1545-5963. doi: 10.1109/TCBB.2014.2372765. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4568560/`.

Anabel Usie, Hiren Karathia, Ivan Teixidó, Rui Alves, and Francesc Solsona. Biblio-MetReS for user-friendly mining of genes and biological processes in scientific documents. *PeerJ*, 2, February 2014. ISSN 2167-8359. doi: 10.7717/peerj.276. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3940481/`.

Yan Wang, Zhiyuan Liu, and Maosong Sun. Incorporating Linguistic Knowledge for Learning Distributed Word Representations. *PLOS ONE*, 10(4):e0118437, April 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118437. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118437`.

Zhehuan Zhao, Zhihao Yang, Hongfei Lin, Jian Wang, and Song Gao. A protein-protein interaction extraction approach based on deep neural network. *International Journal of Data Mining and Bioinformatics*, 15(2):145–164, January 2016. ISSN 1748-5673. doi: 10.1504/IJDMB.2016.076534. URL `http://www.inderscienceonline.com/doi/abs/10.1504/IJDMB.2016.076534`.

Deyu Zhou, Dayou Zhong, and Yulan He. Event trigger identification for biomedical events extraction using domain knowledge. *Bioinformatics*, 30(11):1587–1594, June 2014. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btu061. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu061`.

Alexandre R. Zlotta, Shin Egawa, Dmitry Pushkar, Alexander Govorov, Takahiro Kimura, Masahito Kido, Hiroyuki Takahashi, Cynthia Kuk, Marta Kovylina, Najla Aldaoud, Neil Fleshner, Antonio Finelli, Laurence Klotz, Jenna Sykes, Gina Lockwood, and Theodorus H. van der Kwast. Prevalence of prostate cancer on autopsy: Cross-sectional study on unscreened caucasian and asian men. *Journal of the National Cancer Institute*, 105(14):1050–1058, 2013. doi: 10.1093/jnci/djt151.

Jasmin Šarić, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, 22(6):645–650, 2006. doi: 10.1093/bioinformatics/bti597.

INtERAcT: Interaction Network Inference from Vector Representations of Words

Matteo Manica, Roland Mathis and María Rodríguez Martínez

Supplementary information

1.4 Word distributions

Algorithm 1 Word distributions

```
procedure WordDistributions($\mathcal{W}, CL, KNN$)
    $\mathcal{D} \leftarrow \{\}$                          ▷ Define a matrix to store distributions
    for $w \in \mathcal{W}$ do                             ▷ For all the words
        $D \leftarrow []$                                  ▷ Define a vector to store $w$ neighbors cluster
        $NE \leftarrow KNN(w)$                             ▷ Getting K neighbors
        for $ne \in NE$ do
            append $CL(ne)$ to $D$
        $H \leftarrow histogram(D)$
        append row $H$ to $\mathcal{D}$
    return $\mathcal{D}$
```

1.5 Score analysis

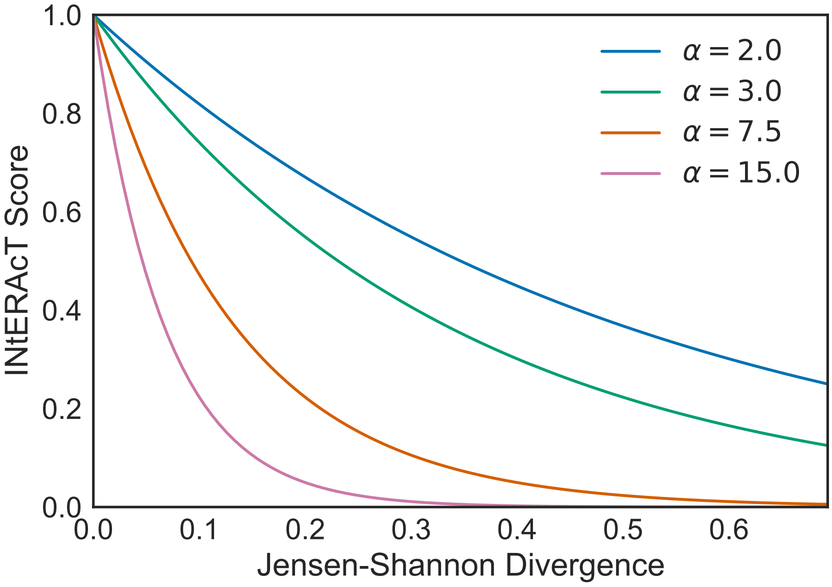


Figure S1: **INtERAcT score analysis.** The curves reported describe how the divergence values are mapped into scores by Eq. 5 setting $\beta = 0.0$ and for different α values. The orange line corresponds to the selected value of $\alpha = 7.5$. Other α values don't map properly the divergence values in a [0,1] interval.

1.6 Prostate–cancer scores

Protein	Protein	Score
MAPK1	MAPK3	0.87
AKT1	MAP2K1	0.83
MAP2K1	MAPK1	0.80
E2F1	FOXO1	0.80
AKT1	MTOR	0.80
CREBBP	EP300	0.78
AKT2	AKT3	0.78
NFKB1	RELA	0.78
MAP2K1	MAPK3	0.77
E2F1	E2F2	0.77
CDKN1A	CDKN1B	0.77
CDKN1B	PTEN	0.77
MAP2K1	RAF1	0.77
CCND1	CCNE1	0.76
AKT1	PIK3R1	0.76
AKT1	PIK3CD	0.75
CCNE1	E2F2	0.75
MDM2	TP53	0.75
E2F2	E2F3	0.75
MAP2K1	PIK3R1	0.75
AKT1	RAF1	0.74
CCNE2	E2F2	0.74
CCND1	CDKN1A	0.74
EGFR	IGF1R	0.74
CDK2	CDKN1A	0.74
CCND1	CDKN1B	0.74
CCNE1	CCNE2	0.74
PDGFB	PDGFD	0.74
CDKN1B	FOXO1	0.73
CCND1	FOXO1	0.73
PIK3CD	PIK3R1	0.73
PIK3R2	RAF1	0.72
CDKN1B	E2F1	0.72
MTOR	PIK3CD	0.72
CDKN1A	E2F1	0.72
MTOR	PIK3R1	0.72
AKT1	AKT3	0.72
KRAS	NRAS	0.72
CDKN1A	FOXO1	0.71
CCNE2	E2F1	0.71
AKT3	RAF1	0.71
AKT1	MAPK1	0.71
CDK2	CDKN1B	0.71
CCND1	E2F2	0.71
PTEN	RB1	0.71
AKT3	MTOR	0.70
E2F1	PTEN	0.70
AKT1	MAPK3	0.70
CCND1	CCNE2	0.70
CDKN1A	TP53	0.70

Table S1: **INtERAcT top–50 scores for KEGG prostate–cancer pathway.** Top–50 interactions predicted from KEGG prostate–cancer pathway using INtERAcT corresponding to the edges of the graph shown in Figure 2a.

1.7 Pubmed Search Queries

	Query on PubMed
KEGG Acute Myeloid Leukemia	”acute myeloid leukemia”
KEGG Bladder Cancer	”bladder cancer”
KEGG Chronic Myeloid Leukemia	”chronic myeloid leukemia”
KEGG Colorectal Cancer	”colorectal cancer”
KEGG Glioma	”glioma”
KEGG Small Cell Lung Cancer	”small cell lung cancer”
KEGG Non Small Cell Lung Cancer	”non small cell lung cancer”
KEGG Pancreatic Cancer	”pancreatic cancer”
KEGG Prostate Cancer	”prostate cancer”
KEGG Renal Cell Carcinoma	”renal cell carcinoma”

Table S2: **PubMed Search queries for KEGG's cancer pathways.** In the Table we report the search query that was used for each KEGG cancer pathway. We used the quotation marks to increase specificity.