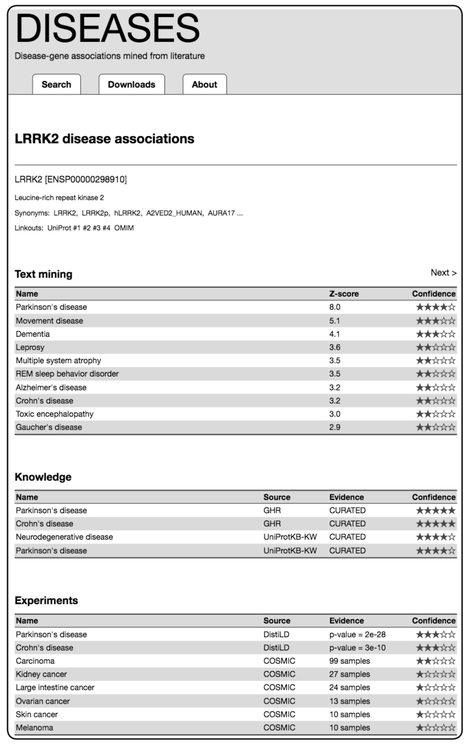
The DISEASES database: improved scoring of disease–gene associations mined from the literature

Alexander Junge and Lars Juhl Jensen
Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark
Contact: alexander.junge@cpr.ku.dk

The Novo Nordisk Foundation
Center for Protein Research

Relevance

Text mining is an established approach to identify associations between biomedical entities. The DISEASES database integrates disease–gene associations found by text mining with expert-curated evidence, cancer mutation data, and genome-wide association studies. Our improved context-aware scoring scheme CoCoScore [3] further improves the text mining pipeline.
DISEASES URL: https://diseases.jensenlab.org/



The DISEASES search interface allows to query genes or diseases.

Detailed evidence viewer for associations found by text mining.

All disease–gene associations, the tagger software, and dictionaries used for text mining are available under open licenses [2]. The tissue expression database TISSUES (https://tissues.jensenlab.org) uses the same text mining approach.

Coverage by evidence channel.

Evidence	Genes	Diseases	Associations
Text mining	15631	4598	478407
Knowledge	2001	735	15231
Experiments	10711	423	89073

Improved text mining via CoCoScore

The DISEASES text mining scheme ignores context when scoring disease–gene pairs. CoCoScore [3] uses textual context to score whether an association is described.
The co-occurrence count C(G,D) of gene G and disease D is defined as:

$$C(G,D) = \sum_{k=1}^{n} \omega_k(G,D) + w_a \delta_{ak}(G,D)$$

Sum over all Medline abstracts

Best score of sentences co-mentioning D and G

Constant score if D and G appear in the same abstract

$$\omega_k(G,D) = \max_{i \in T(k,G,D)} r(i)$$

T(k,G,D) are all sentences in abstract k that co-mention G and D. r(i) is the sentence-level score returned by a fastText-based model [4] (see Figure).

The co-occurrence score S(G,D) is:

$$S(G,D) = C(G,D)^{\alpha} \left(\frac{C(G,D)\ C(\cdot,\cdot)}{C(G,\cdot)\ C(\cdot,D)} \right)^{1-\alpha}$$

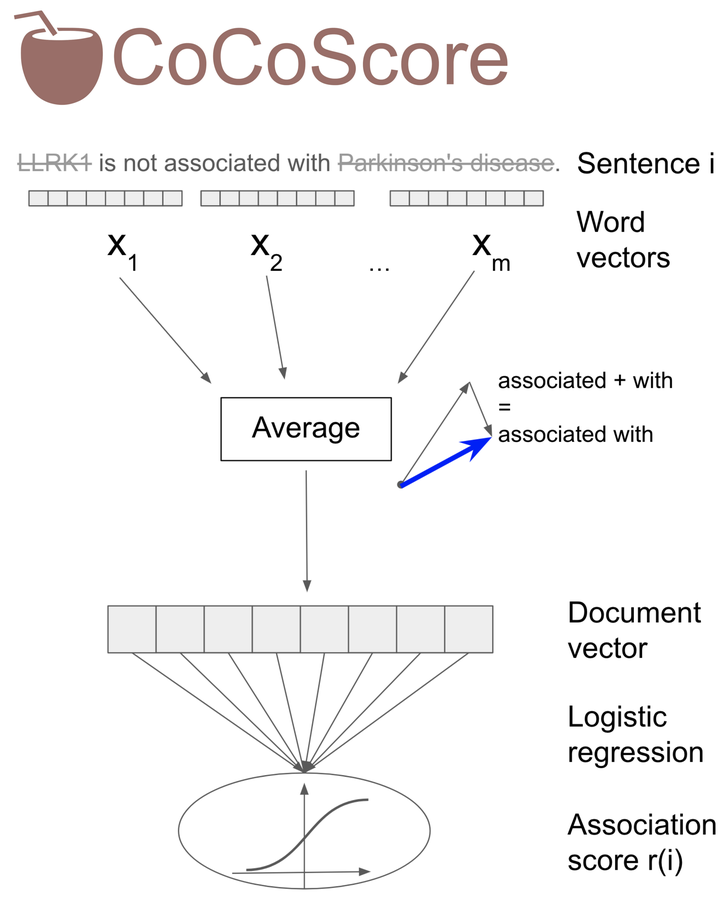
Observed-over-expected ratio

Performance on test dataset derived from curated associations using distant supervision.

	AUROC	AP
CoCoScore	0.97	0.77
DISEASES score	0.96	0.72

AUROC: Area under the ROC curve
AP: Average Precision

CoCoScore



Key features

- DISEASES integrates scored disease–gene associations from text mining, curated knowledge, and experiments
- weekly text mining updates and data downloads
- CoCoScore, our novel context-aware scoring scheme, improves the text mining performance

Funding: Novo Nordisk Foundation, NIH Common Fund

Want to know more?
[1] Pletscher-Frankild *et al.*, Methods (2015)
[2] tagger on Bitbucket: https://goo.gl/hefLqj
[3] CoCoScore on GitHub: https://goo.gl/xCxdjt
[4] Joulin, Grave *et al.*, arXiv (2016)
Visit https://diseases.jensenlab.org/