Introduction to text mining

Lars Juhl Jensen

jensenlab.org

The Novo Nordisk Foundation
Center for Protein Research

FACULTY OF HEALTH SCIENCES
UNIVERSITY OF COPENHAGEN

>10 km

patent literature

grant proposals

too much to read

computer

as smart as a dog

teach it specific tricks


What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!


what they hear
blah blah GINGER blah
blah blah blah blah blah
blah blah GINGER blah
blah blah blah blah...

named entity recognition

dictionary

genes / proteins

diseases / phenotypes

drugs / metabolites

synonyms

alpha-synuclein

SNCA

PARK1

not comprehensive

orthographic variation

spaces and hyphens

alpha-synuclein

alpha synuclein

prefixes and suffixes

SNCA

hSNCA

alternative forms

Parkinson’s disease

Parkinson’s disorder

Parkinson’s syndrome

acronyms

PD

curated blacklist

SDS

software

C++ tagger

>1000 abstracts / second

70–80% recall

80–90% precision

open source

bitbucket.org/larsjuhljensen/tagger/

Docker

hub.docker.com/r/larsjuhljensen/tagger/

web service

tagger.jensenlab.org

EXTRACT

extract.jensenlab.org

primary tumour, establishing the principle of adjuvant therapy[8, 9]. Although this therapy was associated with bone-marrow toxicity, the toxic effects were reversible, whereas the antitumour effects cured patients of their cancer.

The basis for selective effects of these agents against tumour cells versus normal tissue was not apparent from the early laboratory or clinical experiments. It would take 10 years after the initial studies by Farber and colleagues for Michael Osborn and Frank Huennekens to discover, in 1958, that the antifolate drugs specifically inhibi-ted dihydrofolate reductase (DHFR)[10, 11]. Subsequently, Joseph Bertino (Fig. 2), David Goldman, Robert Schimke and Bruce Chabner provided further insight into the mechanisms of methotrexate[12], leading to the model for our current understanding of the pharmacological principles of cancer chemotherapy. The action of methotrexate depends on active transport into cells through the reduced-folate transporter 1 (RFT-1), its conversion to a long-lived intracellular polyglutamate, and its binding to DHFR, which leads to inhibition of the synthesis of thymidylate and

EXTRACT ×
Protein
Chemical compound
Organism
Environment
Tissue
Disease/phenotype
Gene Ontology term

information retrieval

three approaches

ad hoc retrieval

PubMed

text indexing

document similarity

recommendation engines

term vectors

weighting scheme

vector similarity

NER-based retrieval

synonyms

alpha-synuclein

SNCA

ontologies

Parkinson’s disease

neurodegenerative disease

concept indexing

improved recall

too much to read

information extraction

two approaches

natural language processing

syntactic parsing

Gene and protein names

Cue words for entity recognition

Verbs for relation extraction

[nxexpr The expression of
[nxgene the cytochrome genes
[nxpg CYC1 and CYC7]]]
is controlled by
[nxpg HAP1]

Saric et al., *Proceedings of ACL*, 2004

manually crafted rules

a lot of work

machine learning

manually annotated corpus

even more work

co-mentioning

counting

within document

within paragraph

within sentence

weighted count

$$C_{ij} = \sum_{k=1}^{n} \delta_{dijk} w_d + \delta_{pijk} w_p + \delta_{sijk} w_s$$

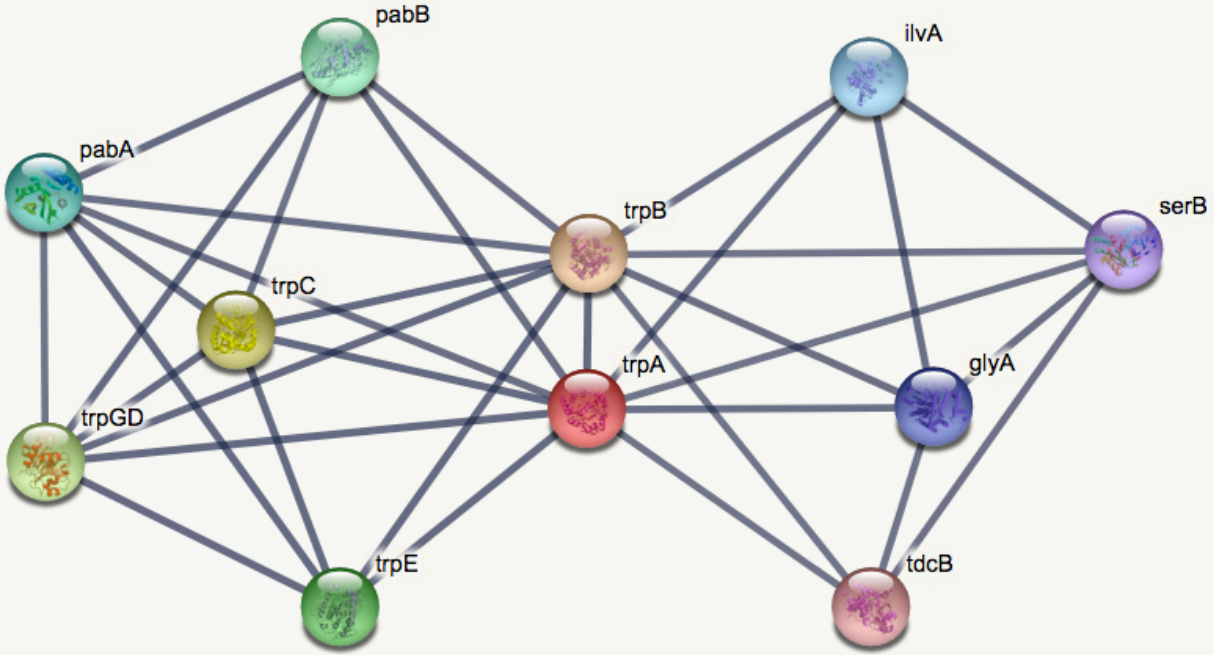
normalization

$$S_{ij} = C_{ij}^{\alpha}\left(\frac{C_{ij}C_{\bullet\bullet}}{C_{i\bullet}C_{\bullet j}}\right)^{1-\alpha}$$

z-scores

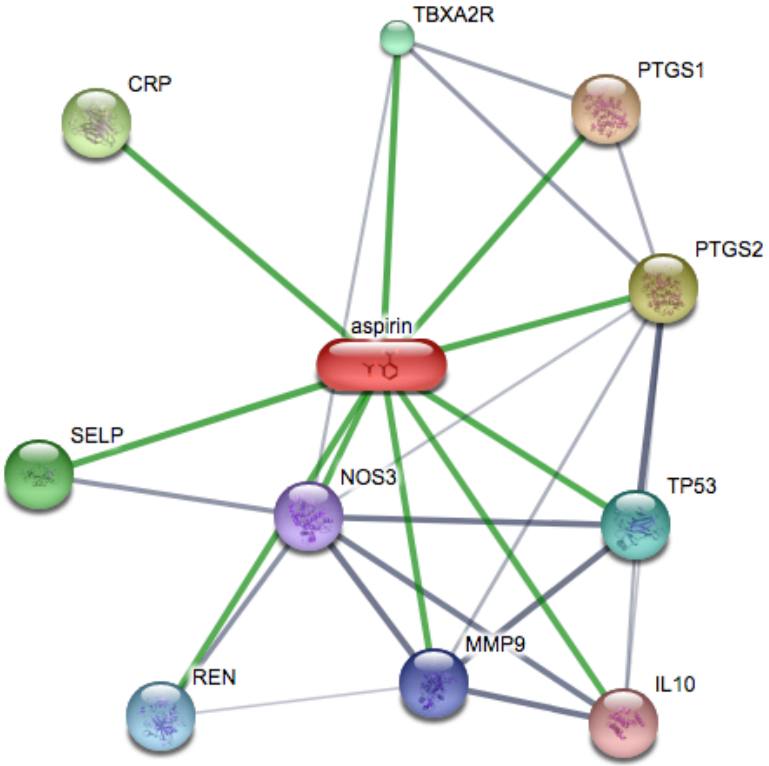
database resources

STRING



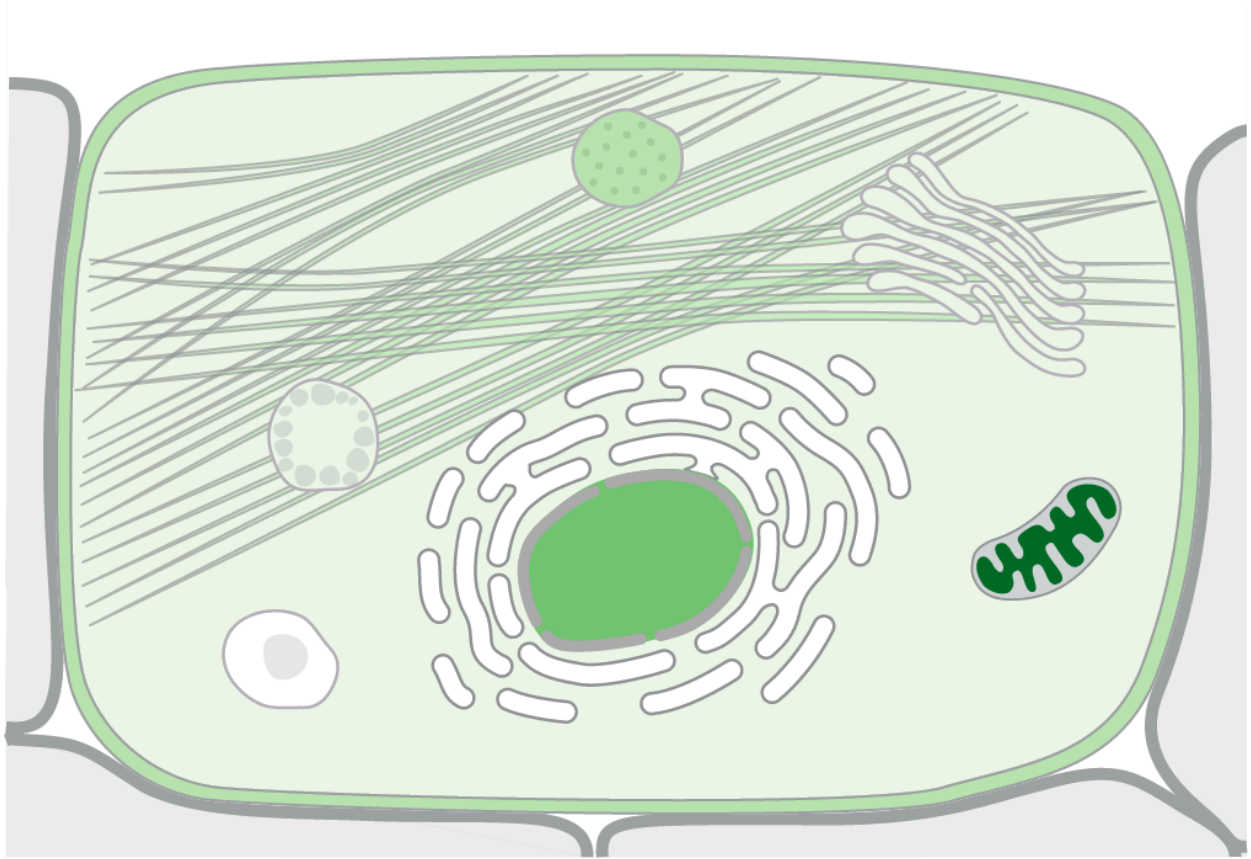
string-db.org

STITCH



stitch-db.org

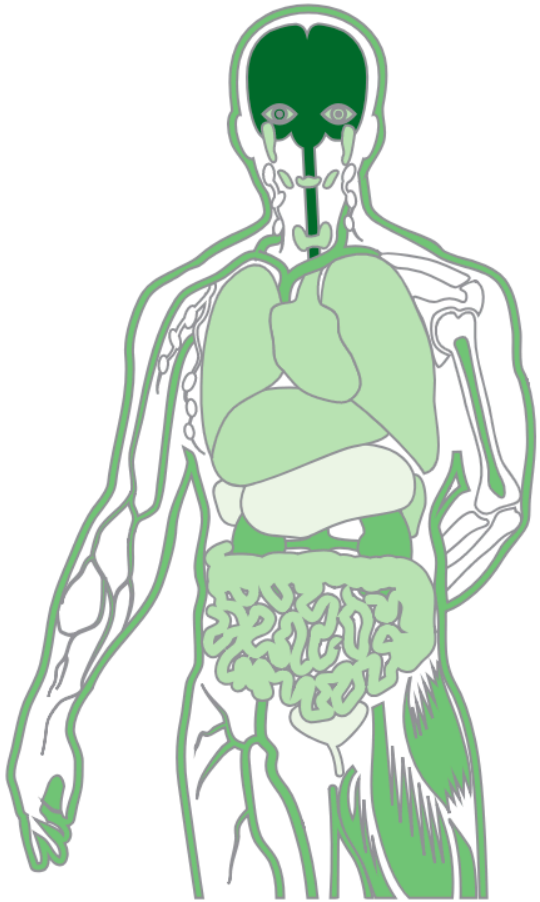
COMPARTMENTS

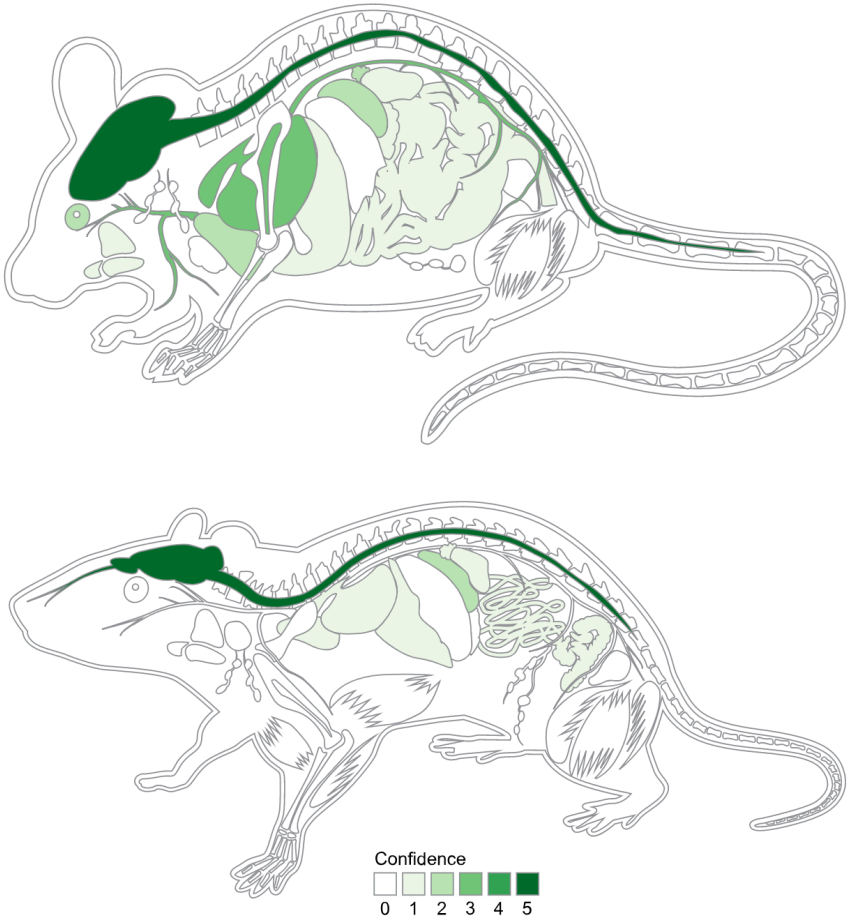


compartments.jensenlab.org

Binder et al., *Database*, 2014

TISSUES





tissues.jensenlab.org

Palasca et al., *Database*, 2018

DISEASES

knowledge discovery

indirect associations

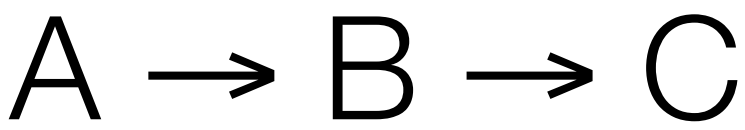
Raynaud’s disease

blood viscosity

fish oil

open discovery

one-node search


A → B → C

generate novel hypotheses

too many hypotheses

closed discovery

two-node search


A → B ← C

find plausible explanations

experimental observation

trend discovery

unknown patterns

known individual facts

enrichment analysis

temporal patterns

conclusions

broadly applicable

keep it simple

free tools