# The Art of Counting

## Scoring and ranking co-occurrences in literature

Lars Juhl Jensen

The Novo Nordisk Foundation
Center for Protein Research

FACULTY OF HEALTH SCIENCES
UNIVERSITY OF COPENHAGEN

1

2

score entity co-occurrences

rank entities for a query

# co-occurrence scoring

# named entity recognition

diseases

genes

# count co-occurrences

what should we count?

within documents

within paragraphs

within sentences

weighted sum

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

famous diseases/genes

observed / expected

$$\frac{C_{ij}C_{\bullet\bullet}}{C_{i\bullet}C_{\bullet j}}$$

single co-occurrences

weighted combination

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

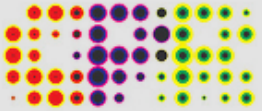
hard to interpret

# z-score transformation

no change to ranking

# DISEASES

Disease-gene associations mined from literature



Search | Downloads | About

## Human genes for idiopathic pulmonary fibrosis

Idiopathic pulmonary fibrosis [DOID:0050156]

A idiopathic interstitial pneumonia which is a distinctive type of chronic fibrosing interstitial pneumonia with thick scarring in the lung creating a honeycomb appearance. The main symptoms start insidiously as shortness of breath on exertion, cough, and diminished stamina. Other common complaints include weight loss and fatigue. The level of oxygen in the blood decreases, and the skin may take on a bluish tinge (called cyanosis) and the ends of the fingers may become thick or club-shape. In most people, symptoms worsen over a period ranging from about 6 months to several years.

Synonyms: idiopathic pulmonary fibrosis, DOID:0050156, FIBROCYSTIC PULMONARY DYSPLASIA, IDIOPATHIC PULMONARY FIBROSIS, FAMILIAL, cryptogenic fibrosing alveolitis ...

### Text mining

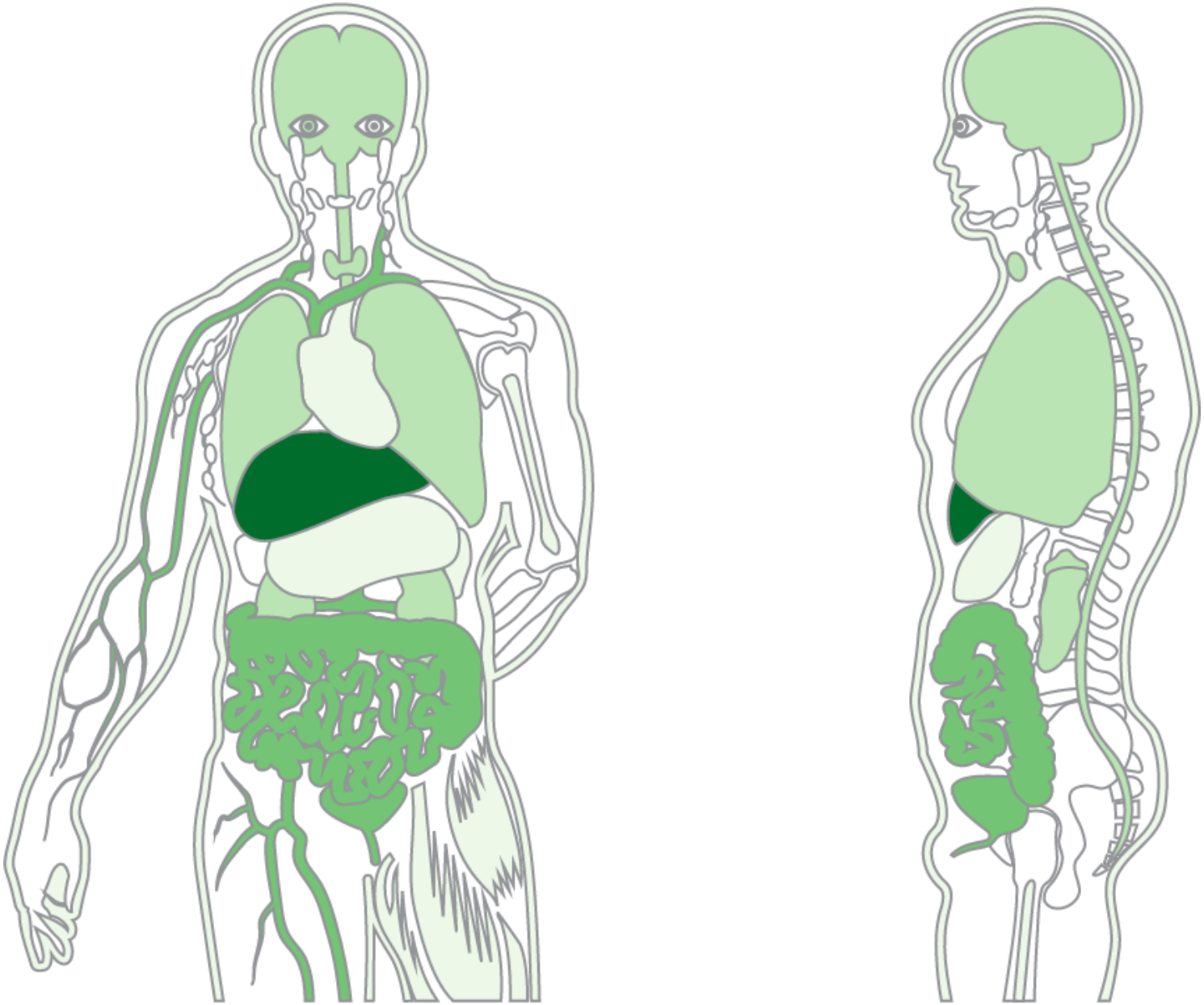
Next >

| Name | Z-score | Confidence |
|---|---|---|
| TGFB1 | 4.1 | ★★★☆☆ |
| SFTPC | 3.9 | ★★☆☆☆ |
| MUC1 | 3.7 | ★★☆☆☆ |
| SFTPD | 3.4 | ★★☆☆☆ |
| ELMOD2 | 3.3 | ★★☆☆☆ |
| FN1 | 3.2 | ★★☆☆☆ |
| TERT | 3.1 | ★★☆☆☆ |
| SFTPA2 | 3.0 | ★★☆☆☆ |
| MMP7 | 2.9 | ★★☆☆☆ |
| CTGF | 2.9 | ★★☆☆☆ |

diseases.jensenlab.org

Frankild et al., *Methods*, 2015

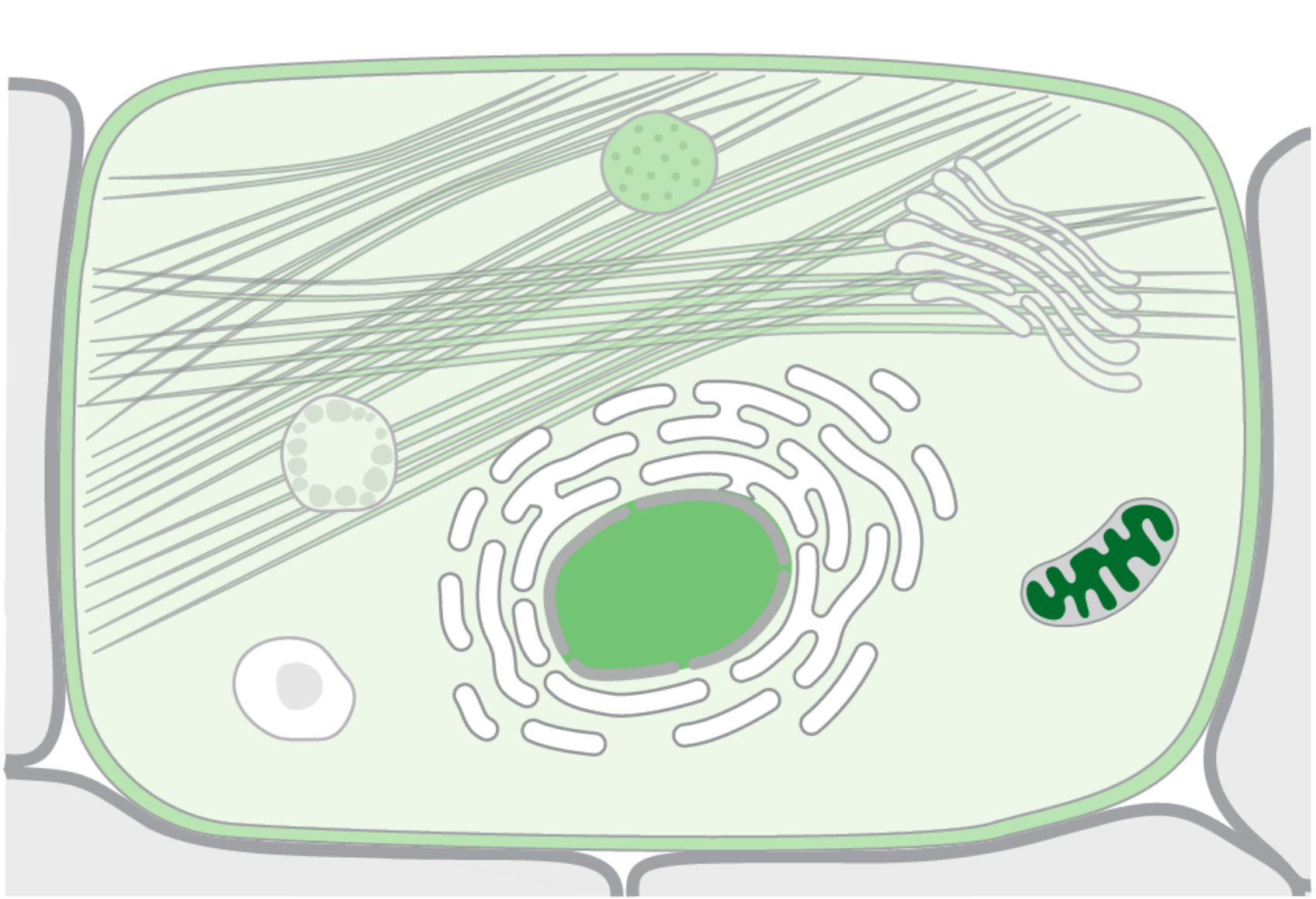
tissue expression



tissues.jensenlab.org

Santos et al., *PeerJ*, 2015

# subcellular localization



compartments.jensenlab.org

Binder et al., *Database*, 2014

# query-based ranking

# rank named entities

*i*

# PubMed query

*j*

query has no position

only count documents

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

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

rank *i* with respect to *j*

# simplified scoring scheme

*j* terms become constant

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

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

no change to ranking

implementation

# PubMed query

# NCBI E-utilities

# PMID list

# relational database

# precompute NER results

precompute all $C_i$.

single SQL query

takes only seconds

# REST API