The Art of Counting
Scoring and ranking co-occurrences in literature

Lars Juhl Jensen
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
\[ \frac{C_{ij}C_{..}}{C_{i..}C_{..j}} \]
single co-occurrences
weighted combination
\[ S_{ij} = C_{ij}^\alpha \left( \frac{C_{ij} C_{..}}{C_{i.} C_{.j}} \right)^{1-\alpha} \]
hard to interpret
z-score transformation
no change to ranking
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

<table>
<thead>
<tr>
<th>Name</th>
<th>Z-score</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFB1</td>
<td>4.1</td>
<td>★★★★★</td>
</tr>
<tr>
<td>SFTPC</td>
<td>3.9</td>
<td>★★★★★</td>
</tr>
<tr>
<td>MUC1</td>
<td>3.7</td>
<td>★★★★★</td>
</tr>
<tr>
<td>SFTP4D</td>
<td>3.4</td>
<td>★★★★★</td>
</tr>
<tr>
<td>ELMOD2</td>
<td>3.3</td>
<td>★★★★★</td>
</tr>
<tr>
<td>FN1</td>
<td>3.2</td>
<td>★★★★★</td>
</tr>
<tr>
<td>TERT</td>
<td>3.1</td>
<td>★★★★★</td>
</tr>
<tr>
<td>SFTP4A2</td>
<td>3.0</td>
<td>★★★★★</td>
</tr>
<tr>
<td>MMP7</td>
<td>2.9</td>
<td>★★★★★</td>
</tr>
<tr>
<td>CTGF</td>
<td>2.9</td>
<td>★★★★★</td>
</tr>
</tbody>
</table>
tissue expression
subcellular localization
query-based ranking
rank named entities
PubMed query
query has no position
only count documents
\[ C_{ij} = \sum_{k=1}^{n} \delta_{dijk} w_d + \delta_{pijk} w_p + \delta_{sijk} w_s \]
\[ C_{ij} = \sum_{k=1}^{n} \sum_{dijk}^{W_{ij}} \]
rank \( i \) with respect to \( j \)
simplified scoring scheme
$j$ terms become constant
\[ S_{ij} = C_{ij}^\alpha \left( \frac{C_{ij} C_{..}}{C_{i.} C_{.j}} \right)^{1-\alpha} \]
\[ S_{ij} = C_{ij}^\alpha \left( \frac{C_{ij}}{C_{i..}C_{.j}} \right)^{1-\alpha} \]
no change to ranking
implementation
PubMed query
NCBI E-utils
PMID list
relational database
precompute NER results
precompute all $C_i$. 
single SQL query
takes only seconds
REST API