Beegle: A Google-like Tool for Disease-Gene Annotation based on Literature Mining

Sarah ElShal
Amin Davani
Jesse Davis
Yves Moreau

17/05/2014
WHY – (1) Disease-Gene Annotation

What if we could save an elder from the complications of Alzheimer’s disease...

What if we could find a treatment to a child with a rare genetic disorder...

What if we could offer physicians extra time and help more patients...
WHY – (2) Literature Mining

More than 23 million citations for biomedical literature.
WHAT – Google-like Tool

Sarah ElShal - KU Leuven
Alzheimer's disease (AD) is the most common form of dementia. To date, several genes have been identified as the cause of AD, including PSEN1, PSEN2, and APP. The association between APOE and late-onset AD has also been reported. We here used a bench top next-generation sequencer, which uses an integrated semiconductor device, detects hydrogen ions, and operates at a high-speed using nonoptical technology.

PMID: 24566006

1- Common Abstracts

Alzheimer's disease (AD) is a neurodegenerative disorder that occurs due to progressive deposition of amyloid β-protein (Aβ) in the brain. Stable conformations of solvated Aβ1-42 protein were predicted by molecular dynamics (MD) simulation using the OPLSAA force field. The seven residue peptide (Lys-Leu-Val-Phe-Phe-Ala-Glu) Aβ16-22 associated with AD was studied and reported in this paper.

PMID: 23320570

These data show that the modified Y-maze is sensitive to age-related deficits, but not additional memory deficits due to amyloid pathology in APP/PSEN1 mice. They also suggest improvements in short-term spatial memory were not due to changes in the neuropathological features of AD or monoamine signaling.

PMID: 23320570

2- Common Concepts
(1) Common Abstracts

Caspase-6 is an effector caspase that has not been investigated thoroughly despite the fact that Caspase-6 is strongly activated in Alzheimer disease brains...

PMID: 24265764

Less than 5% of all Alzheimer's disease cases are familial in nature, i.e. caused by mutations in APP, PSEN1 or PSEN2...

PMID: 24101602

Presenilin 1 (PSEN1) gene mutations deterministic for Alzheimer's disease (AD) are associated with marked heterogeneity in clinical phenotype...

PMID: 23948899

Upon PSEN1/2 re-introduction, this active epigenetic state was replaced by a MeCP2-containing repressive state and reduced Neurotrypsin expression...

PMID: 24145027

Sarah ElShal - KU Leuven

### Results

<table>
<thead>
<tr>
<th>Alzheimer's Disease</th>
<th>Not Alzheimer's Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSEN1</td>
<td>X</td>
</tr>
<tr>
<td>Not PSEN1</td>
<td>K</td>
</tr>
</tbody>
</table>

\[
\text{score} = p-value = \frac{K \binom{M-K}{N-X}}{\binom{M}{N}}
\]
(2) **Common Concepts**

**Alzheimer’s Disease**

<table>
<thead>
<tr>
<th>Alzheimer's Disease</th>
<th>Amyloid</th>
<th>Presenile dementia</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.034075</td>
<td>0.074095</td>
<td>0.051059</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alzheimer's Disease</th>
<th>Amyloid</th>
<th>Presenile dementia</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.041778</td>
<td>0.057728</td>
<td>0.021259</td>
</tr>
</tbody>
</table>

**tf-idf scores**

**PSEN1**

**MetaMap Portal**

**UMLS®**

**WHY**

**WHAT**

**HOW**

**RESULTS**

\[
\text{score} = \frac{(tf \times \log(idf)_{\text{query}} \times tf \times \log(idf)_{\text{gene}})}{\|tf \times \log(idf)_{\text{query}}\| \|tf \times \log(idf)_{\text{gene}}\|}
\]
Results – **OMIM Benchmark (1)**

### morbidmap

~6000 entries

- Alzheimer disease 1, familial, 104300 (3) | **APP**, AAA, CVAP, AD1 | 104760 | 21q21.3
- Alzheimer disease 17 (2) | AD17 | 615080 | 6p21.2
- Alzheimer disease 6, 104300 (2) | AD6 | 605526 | 10q24
- Alzheimer disease 8, 104300 (2) | AD8 | 607116 | 20p
- Alzheimer disease, type 3, 607822 (3) | **PSEN1**, AD3 | 104311 | 14q24.2
- Alzheimer disease, type 3, with spastic paraparesis and apraxia, 607822 (3) | **PSEN1**, AD3 | 104311 | 14q24.2
- Alzheimer disease, type 3, with spastic paraparesis and unusual plaques, 607822 (3) | **PSEN1**, AD3 | 104311 | 14q24.2
- Alzheimer disease-10, 104300 (2) | AD10 | 609636 | 7q36
- Alzheimer disease-11 (2) | AD11 | 609790 | 9p22.1
- Alzheimer disease-2, 104310 (3) | **APOE**, AD2, LPG, LDLCQ5 | 107741 | 19q13.32
- Alzheimer disease-4, 606889 (3) | **PSEN2**, AD4, STM2, CMD1V | 600759 | 1q42.13
- Alzheimer disease-5, 104300 (2) | AD5 | 602096 | 12p11.23-q13.12
- Alzheimer disease-7 (2) | AD7 | 606187 | 10p13

... - Strong mapping evidence “3”
- Official symbols “HUGO”
- Combined entries “Levenshtein distance”
- Reliable entries “>=3 genes”

### benchmark

~300 entries

...  
14 Alzheimer disease [SORL1, HFE, A2M, BLMH, APP, ACE, PACIP1, PSEN1, APOE, PSEN2, MPO, NOS3, APBB2, PLAU] ...
Results – *OMIM Benchmark (2)*

1# Common Abstracts

2# Common Concepts

3# Combined

<table>
<thead>
<tr>
<th>Alzheimer’s Disease Genes</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SORL1</td>
<td>5</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td>PSEN2</td>
<td>7</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>PSEN1</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

~80% recall in top 100
**Demo**

---

**Alzheimer disease**

**AMPA**
- Amyloid beta (Aβ) precursor protein
- Amyloid beta (Aβ) peptides, which are generated from amyloid precursor protein (APP), are thought to play a major role in the pathogenesis of Alzheimer's disease (AD). This study investigated the anti-amyloidogenic effects of the ethanolic extract of *Melia fructus* (ID1201) using human embryonic kidney 293 cells with stably expressed human wild-type or Swedish mutant APP695 and β-secretase 1. ID1201 treatment enhanced the non-amyloidogenic metabolism of APP; increases in soluble APPβ levels and decreases in soluble APPα and Aβ levels resulted from the α-secretase activation through the phosphatidylinositol 3-kinase (PI3K)/Akt pathway...  
  **PMID:** 18568886

**MAPT**
- Microtubule-associated protein tau
- Alzheimer's disease is the most common form of dementia. Abnormal hyperphosphorylation of microtubule-associated protein tau (MAPT) is one of the hallmarks of Alzheimer's disease and related tau pathies. CDK5 and GSK3β are the two main protein kinases that have an important role in the abnormal hyperphosphorylation of MAPT which leads to Alzheimer's disease...  
  **PMID:** 24402738

**BACE1**
- Beta-site APP-cleaving enzyme 1
- Environmental exposure to lead (Pb) early in life results in a latent upregulation of genes and products associated with Alzheimer's disease (AD), particularly the plaque forming protein amyloid beta (Aβ). Furthermore, animals exposed to Pb as infants develop cognitive decline and memory impairments in old age. Studies from our lab demonstrated that tolfenamic acid lowers the levels of the amyloid β precursor protein (APP) and its aggregative cleavage product Aβ by inducing the degradation of the transcription factor specificity protein 1 (Sp1)...  
  **PMID:** 24402738
Future Work

1. Enhanced Vocabulary

2. Identification + Discovery

3. Free-text Queries
Thank you!