Intelligent Bibliometrics for Gene-Disease Association Analysis and Prediction

Mengjia Wu, Yi Zhang
Australian Institute of Artificial Intelligence
University of Technology Sydney
2020.8.1
1. Research Motivation

Genetic Analysis for Disease: occurrence, diagnosis and treatment

Data-driven Disease-Gene Association Prediction:
- Curated Databases – limited knowledge within established frameworks
- Literature Based Discovery (LBD) – the requirement of expert knowledge

- Propose an adaptable and automatic LBD approach for the following tasks:
  1. How to identify the crucial genetic entities for a specific disease.
  2. How to predict emerging genetic factors for the target disease.
2. Methodology Framework

Stage 1
Data Collection and Pre-processing

Stage 2
Bioentity2Vec Training and Network Construction

Stage 3
Network Analytics
2. Methodology Framework

- Heterogenous Network Construction

- Disease: target disease, symptoms, risk factors, complications etc.
- Chemical: chemical elements, compounds, drugs etc.
- Gene: refers to a certain segment of nucleotides on chromosome;
- Genetic variant: gene mutation, protein mutation and single nucleotide polymorphism (SNP)
2. Methodology Framework

- Network Analytics – Centrality Measurement

Degree Centrality ($DC$)

$DC(A) = \frac{\text{The degree of } A}{\text{Num of nodes} - 1}$

For node A, $DC = \frac{3}{5} = 0.6$
2. Methodology Framework

- Network Analytics – Centrality Measurement

Closeness Centrality ($CC$)

$$CC(A) = \frac{\text{Num of nodes} - 1}{\text{the sum of topological distances of } A \text{ to other nodes}}$$

For node A, $$CC = \frac{5}{1+1+1+2+2} = 0.714$$
2. Methodology Framework

• Network Analytics – Centrality Measurement

![Diagram]

Betweenness Centrality ($BC$)

$$BC(V_t^m) = \frac{\sum_{all\ pairs} num\ of\ the\ shortest\ paths\ pass\ A}{Total\ num\ of\ the\ shortest\ paths\ the\ num\ of\ node\ pairs}$$

For node A, $$BC = \frac{1 + \ldots + \ldots}{(5*4)/2}$$
2. Methodology Framework

- Centrality Integration: Non-dominating sorting\cite{2}

<table>
<thead>
<tr>
<th></th>
<th>Degree Centrality</th>
<th>Closeness Centrality</th>
<th>Betweenness Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node A</td>
<td>0.8</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Node B</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Node C</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Node D</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Node E</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- Objective: Comprehensively identify dominant nodes with 3 prior values for all the centralities

2. Methodology Framework

• Network Analytics – Link Prediction

- Common neighbor-based Assumption: If two unconnected nodes share common neighbor(s), there is possibility that an edge will emerge between them.
2. Methodology Framework

- Link Prediction - Resource Allocation\[^3, 4\]

\[
\text{Resource Allocation Index (B, C)} = \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{1}{|\Gamma(w)|} \left( \frac{1}{2} + \frac{1}{3} \right) = 0.833
\]

\[
\text{Resource Allocation Index (B, C) (weighted version)} = \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{E(w,B) + E(w,C)}{\sum_{v \in \Gamma(w)} E(w,v)}
\]

---


2. Methodology Framework

- **Bioentity2Vec Model Training**
  - …Plasma big endothelin-1 predicts atrial fibrillation … late gadolinium enhancement…of AF and fibrosis…
  - Skip-Gram Algorithm\[1\]
  - Gene
  - Disease
  - Chemical
  - Entity Window size = 5

- **Semantic Similarity ("AF", “ET-1”) = Cosine Similarity (\(\overrightarrow{AF}, \overrightarrow{ET-1}\))**

2. Methodology Framework

• Bioentity2Vec & Resource Allocation Incorporation

Proposed Semantic-Enhanced Resource Allocation Index:

\[ R_{(B,C)} = \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{CF(B, w) S_{B,w} + CF(w, C) S_{w,C}}{\sum_{v \in \Gamma(w)} CF(v, w) S(v,w)} \]

*CF*(B, w) is the co-occurring frequency of entity B and entity w, *S*_{B,w} represents the semantic similarity between entities B and w.

Output: a ranking list of genetic factors
3. Case Study

- Data Collection and Entity Extraction
- PubMed database

“("Atrial Fibrillation"[Mesh] AND Humans[Mesh])”
Search Date: 2020/04/28

Record Num: 54,219
3. Case Study

- Entity Extraction and Pre-processing

Entity Extraction using Pubtator

- Diseases & Chemicals
  - MeSH Dictionary
- Genes
  - NCBI Gene Dictionary
- Genetic Variants
  - dbSNP Database

Remove Isolated Nodes

5,838 nodes

6,318 biomedical entities

6,318 biomedical entities
Cerebrovascular events, bleeding complications and device related thrombi in atrial fibrillation patients with chronic kidney disease and left atrial appendage closure with the WATCHMAN device

PMID31092201
LUANI B, GENZ C ... RAUWOLF T • BMC CARDIOVASC DISORD. 2019 MAY 15 • 2019

Background

Impaired renal function increases the bleeding risk, leading to a conservative prescription and frequent discontinuation of oral anticoagulation in atrial fibrillation patients with chronic kidney disease (CKD). Interventional left atrial appendage closure (LAAC) might be an alternative therapeutic strategy for these patients.
3. Case Study

- Centrality Measurement

![Graph showing centrality measurements with labels for KCNA5, SCN5A, AGT, GJA5, F10, ACE, TNF, F2, IL6, NPPB, and CRP. The graph is a 3D scatter plot with axes for Degree Centrality, Closeness Centrality, and Betweenness Centrality.]
### 3. Case Study

- **Centrality Measurement - Gene**

<table>
<thead>
<tr>
<th>Top 20 Results by Non-dominating Sorting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial Fibrillation; Stroke; Heart Failure; Hypertension; Hemorrhage; Diabetes Mellitus; Fibrosis; Myocardial Infarction; Cerebral Infarction; Ischemia; Thromboembolism; Death; Thrombosis; Inflammation; Coronary Artery Disease; Tachycardia; Ventricular Fibrillation; Tachycardia, Supraventricular; Neoplasms; Atrioventricular Block</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warfarin; Calcium; Amiodarone; Potassium; Digoxin; Ethanol; Verapamil; Sodium; Oxygen; Quinidine; Aspirin; Vitamin K; Glucose; Cholesterol; apixaban; Sotalol; Nitrogen; Magnesium; Heparin; Propafenone</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chemical</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRP; F2; ACE; IL6; AGT; F10; SCN5A; NPPB; KCNA5; PITX2; FGB; GJA5; TNNI3; INS; TNF; TGFβ1; VWF; KCNQ1; SERPINE1; AGTR1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2200733; rs6795970; rs2106261; rs2108622; rs3789678; rs13376333; rs17042171; rs1805127; rs7539020; rs11568023; rs10033464; rs3807989; rs7193343; rs3918242; rs3825214; rs16899974; rs699; rs7164883; rs6584555; rs10824026</td>
</tr>
</tbody>
</table>
3. Case Study

• Link Prediction Validation

Roll Back the dataset by 5 years
## 3. Case Study

### Validation Results

<table>
<thead>
<tr>
<th></th>
<th>Resource Allocation</th>
<th>Weighted Resource Allocation</th>
<th>Modified Resource Allocation (Purposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top $k$ Recall</td>
<td>0.245</td>
<td>0.208</td>
<td>0.283</td>
</tr>
<tr>
<td>Top 100 Recall</td>
<td>0.434</td>
<td>0.396</td>
<td>0.472</td>
</tr>
<tr>
<td>Top 200 Recall</td>
<td>0.604</td>
<td>0.642</td>
<td>0.736</td>
</tr>
</tbody>
</table>

# $k$ refers to the number of edges that were removed for node AF, in this experiment $k = 53$. 
4. Limitations and Future Directions

Limitations:
• Negative associations collected when using co-occurrence
• The genetic research of AF is still at an early stage, some associations between AF and genes haven’t been revealed yet

Future Study:
• Employ Sentiment analysis to exclude those negative associations
• Modify the entity extraction rules
• Involve the identified crucial genetic factors to improve predicting performance
Thank you!

Email address: Mengjia.wu@student.uts.edu.au