Batyр Charyyyev
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Outline

• Introduction

• Methodology
  ➢ Retrieval of abstracts
  ➢ Preprocess abstracts
  ➢ Extraction and characterization of relations
  ➢ Generating network

• Evaluation

• Conclusion
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Introduction

• Problem
  - 24,000 genes according to Human Genome Project [1]
  - Enormous amount of data
  - Difficult to analyze and interpret
Introduction

• **Goal**
  - Extract information
  - Provide complete and comprehensive view

• **Expected outcome**
  - Improvement in result of extraction.
  - Find different characteristics of genes from network
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Introduction

Retrieve
- Check for aliases
- Parshehub
- Pubmed

Preprocessing data
- Threshold scores

Generating network
- Gephi
- Generate network properties

Extraction and characterization
- Check for synonyms
- Rule based
- Weight edges
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Data set

- Gene: HUGO Gene Nomenclature Committee (HGNC)
- Relations: PubMed
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

**Hugo Gene Nomenclature Committee (HGNC)**

- Name and symbol
- Locus group and type
- Alias symbol and name
- Date approved and modified
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Retrieval of abstracts from Pubmed with ParseHub
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Retrieval of abstracts from Pubmed with ParseHub

- Aliases of gene from HGNC
- Time limit

- **Advantage**: Retrieves all papers
- **Disadvantage**: Time required
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Preprocessing data

- **Problem:** Less amount of related work
  - Irrelevant data
  - Overfitting the problem

- **Solution:** Apply some thresholding
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Preprocessing data

- Apply threshold

$$\text{Score } (c_1, c_2) = \frac{a_j}{N_j} \cdot c_1 + r_j \cdot c_2$$

- $a_j$: number of occurrence of gene in abstract
- $N_j$: number of words in abstract
- $r_j$: rank of paper determined by PubMed
- $c_1, c_2$: constants
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Extraction and characterization

• Identify entities like genes, biological concepts

  ➢ **Problem:** Different entities, with same meaning

  ➢ **Solution:** Clean from aliases.

• Rule based

• Weight edges
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Extraction and characterization

- Extraction of biological interaction network from scientific literature [2]
- Content-rich biological network constructed by mining Pubmed abstracts [1]
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Extraction and characterization

\[ w = a \cdot S_s + (1 - a) \cdot S_a + a \cdot R \cdot S_s + \frac{1}{N_b + 1} \cdot (C + T) \]  

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_a )</td>
<td>Number of abstracts containing two related genes.</td>
</tr>
<tr>
<td>( S_s )</td>
<td>Number of sentence containing two related genes.</td>
</tr>
<tr>
<td>( a )</td>
<td>Constant to determine significance of being in same sentence.</td>
</tr>
<tr>
<td>( R )</td>
<td>Coefficient of regulatory word (Suppresses, activates).</td>
</tr>
<tr>
<td>( N_b )</td>
<td>Number of sentence between two related genes.</td>
</tr>
<tr>
<td>( C )</td>
<td>Constant to determine significance of being close.</td>
</tr>
<tr>
<td>( T )</td>
<td>Coefficient of transition word (However, furthermore) between sentences.</td>
</tr>
</tbody>
</table>
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Extraction and characterization

\[ w = a \cdot S_s + (1-a) \cdot S_a + a \cdot R \cdot S_s + \frac{1}{N_b + 1} \cdot (C + T) \]

- Being in same sentence also means being in same abstract
- Being in same sentence and having regulatory word (suppress, activates)
- How close two sentence?
- Is there transition word between two sentence?
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Generating network

• **Tools**: Gephi

• **Properties**:
  - Clustering
  - Communities
  - Assortative property
  - Evolution of network
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Generating network

- **Assortative:**
  - Locus group
  - Gene family
  - Location

- **Network evolution:**
  - Date of publication
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Evaluation

- Add random abstracts with random ranking
- Filter abstracts based on score
- Evaluate the result based on precision
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Evaluation

• Chilibot: Content-rich biological network constructed by mining PubMed abstracts
  ➢ Biological concepts
  ➢ Weight, number of abstracts

• http://www.chilibot.net/
Gene Relation Network Constructed by Mining Biomedical Literature Abstracts.

Evaluation:

idp1 idh2 mitochondria DNA petite
Conclusion

• **Goal:** Provide complete and comprehensive view of data

• **Objectives:** Retrieve and extract information and generate network

• **Difference:**
  - Check relevance of paper
  - Weight of relations
  - Focusing only to genes

• **Data Set:** PubMed, HGNC

• **Tools:** ParseHub, Gephi
References


Questions