KATZLGO: Large-scale Prediction of LncRNA Functions by Using the KATZ Measure Based on Multiple Networks
Contents

1. Introduction
2. Materials
3. Methods
4. Results
Introduction

- LncRNA can execute a broad range of important functions
- Only a small number of lncRNAs have been functionally well-characterized
- Besides experiments, several prediction methods of lncRNA functions have been developed
- Most of the methods are based on gene expression and local information
Materials

IncRNA-Protein associations
Co-expression data:
COXPRESdb
ArrayExpress and GEO
Interaction data:
Npinter 3.0

IncRNA co-expression similarity
LncRNA expression profiles:
NONCODE2016

Protein-Protein Interactions
Protein-protein interaction:
STRING
Methods (pipeline)

- IncRNA co-expression
- IncRNA-protein association
- Protein-protein interaction network

\[ A = \begin{bmatrix} L & LP \\ LP^T & P \end{bmatrix} \]

\[ S_{Katz}(i, j) = \sum_{i=1}^{\infty} \beta^i A^i(i, j) \]

- IncRNA-IncRNA similarities
- IncRNA-protein associations
- Protein-protein interactions

Similarity score matrix:

- GO1, GO2, ..., GO1
- GO3, GO2, ..., GO2
- GO7, GO5, ..., GO3

GO prediction results:

- GO3, GO5, ..., GO3
- GO7, GO1, ..., GO2
- GO5, GO3, ..., GO1

The global heterogeneous network
**Methods (KATZ measure)**

- Suppose the global network is looked upon as an undirected graph denoted as an adjacency matrix $A$.
- The similarity between node $i$ and node $j$ is given by
  \[ S_{Katz}(i, j) = \sum_{l=1}^{\infty} \beta^l A^l(i, j) = (I - \beta)^{-1} - I \]

\[
A = \begin{bmatrix}
L & LP \\
LPT & P
\end{bmatrix}
\]

- Assuming $k = 3$, the portion of the KATZ similarity matrix $S_{katz}$ corresponding to scores between lncRNA nodes and protein nodes, denoted as $S_{LP}$, is written as:
  \[
  S_{LP} = \beta LP + \beta^2 (L \ast LP + LP \ast P) + \beta^3 (LP \ast LPT \ast LP + L^2 \ast LP + L \ast LP \ast P + LP \ast P^2)
  \]
The probability $P_i(T_i)$ for each GO term $T_i$ assigned to the query lncRNA $I$ is the sum of weights of neighboring protein-coding gene with the annotation $T_i$

$$P_i(T_i) = \sum_{i=1}^{N} \frac{S_{lp}(i)}{\sum_{j=1}^{N} S_{lp}(j)} \cdot \text{Ind}(T_i)$$

$\text{Ind}(T_i)$ is an indicator function defined as:

$$\text{Ind}(T_i) = \begin{cases} 
1 & \text{if } I_i \text{ has the annotation } T_i \\
0 & \text{otherwise}
\end{cases}$$
Results (Benchmarks)

- **IncRNA2GO-55:**
  manually annotated 55 lncRNA with 129 GO terms

- **Protein2GO-825:**
  the proteins with GO annotations extracted from UniProt-GOA version 201604 excluding those in UniProt-GOA version 201010
Results

- The effects of protein interaction data

![Graph showing precision and recall](image)

- Protein2GO-825 dataset for Molecular Function
- lncRNA2GO-55 dataset for Molecular Function
Results

- **Performance of KATZLGO**

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>Fmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>KATZLGO</td>
<td>0.583</td>
<td>0.176</td>
<td>0.325</td>
</tr>
<tr>
<td>lnc-GFP</td>
<td>0.223</td>
<td>0.228</td>
<td>0.225</td>
</tr>
<tr>
<td>LncRNA2Function</td>
<td>0.531</td>
<td>0.095</td>
<td>0.161</td>
</tr>
</tbody>
</table>

Performance comparison with the methods of lnc-GFP and LncRNA2Function

The numbers of 55 lncRNAs annotated correctly based on KATZLGO vs. the other two methods (lnc-GFP and LncRNA2Function)