Introduction

Identifying which genes are involved in particular biological processes is relevant to understand the structure and function of a genome. A number of techniques have been proposed that aim to annotate genes, i.e., identify unknown biological associations between biological processes and genes. The ultimate goal of these techniques is to narrow down the search for promising candidates to carry out further studies through in-vivo experiments.

Our work presents an approach for in-silico prediction of functional gene annotations. It uses existing knowledge body of gene annotations of a given genome and the topological properties of its gene co-expression network, to train a supervised machine learning model that is designed to discover unknown annotations. The approach is applied to Oryza sativa japonica (a variety of rice).

Methodology

Gene Co-expression Network

Gene co-expression networks are represented as undirected graphs where each vertex identifies a gene and an edge the level of co-expression between two genes.

The information is taken from the ATTED-II database [2]. The gene co-expression network \( G = (V, E, w) \) comprises 19,665 vertices (genes) and 553,125 edges. The weight function \( w : E \rightarrow \mathbb{R}_{>0} \) measures the co-expression between any pair of genes.

Gene Functional Annotations

Each gene is associated with the collection of functional annotations (biological processes) to which it is related (e.g., through in-vivo experiments).

The annotation information is taken from the RAP-DB [3] database, a comprehensive set of gene annotations for the genome of rice. There are 633 annotations for biological processes (i.e., pathways to which a gene contributes). It is important to note that genes may be associated to several annotations.

Topological Properties

Given the co-expression network \( G = (V, E, w) \), properties of its network structure are computed for gene annotation prediction. Topological measures considered for each gene \( u \) are the following:

- degree: number of edges incident to \( u \);
- eccentricity: maximum shortest distance from \( u \) to any vertex in its connected component;
- clustering coefficient: ratio between the number of triangles (3-loops) that pass through \( u \) and the maximum number of 3-loops that could pass through it;
- closeness centrality: the reciprocal of the average shortest path length from \( u \);
- betweenness centrality: the amount of control that \( u \) has over the interactions of other nodes in the network;
- neighborhood connectivity: the average connectivity of all neighbors of \( u \);
- topological coefficient: the extent to which \( u \) shares neighbors with other nodes.

Supervised Training

Two models are trained per biological function for predicting gene annotations. Namely, one in which the topological measures of \( G \) are used and another one in which they are not. The dataset summarizes data for 19,665 genes, 615 annotations, and 7 topological measures.

The dataset is heavily imbalanced since 77% of annotations are related to less than 10 genes each one. Only annotations associated with at least 10 genes are considered for prediction (141). The Synthetic Minority Over-sampling Technique (SMOTE) is used to over-sample the minority class.

The supervised machine learning technique XGBoost is used for annotation prediction [1]. This technique is a Python implementation of gradient boosted decision trees.

Results

Figure 2 shows that the model trained with additional information of the topological measures can be more reliable in some cases.

Table 1: Number of genes most frequently annotated as false positives by the model trained with topological measures.

<table>
<thead>
<tr>
<th>ID</th>
<th>Biological process</th>
<th># Genes</th>
<th>Max FP</th>
<th># FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0006807</td>
<td>nitrogen compound metabolic process</td>
<td>15</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>0006289</td>
<td>nucleotide-excision repair</td>
<td>20</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>0006397</td>
<td>mRNA processing</td>
<td>17</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>0007017</td>
<td>microtubule-based process</td>
<td>18</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>0006506</td>
<td>calcium ion transmembrane transport</td>
<td>10</td>
<td>36</td>
<td>1</td>
</tr>
</tbody>
</table>

Acknowledgment

This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), sponsored within the Colombian Scientific Ecosystem by The World Bank, Colciencias, Icetex, the Colombian Ministry of Education and the Colombian Ministry of Industry and Tourism under Grant FP44842-217-2018.

References

