From Gene Expression modelling to CoRegulation networks for Arabidopsis

NETBIO
Thursday 18th September 2014
A huge ‘orphan of function’ gene space:

- 20% to 40% of the predicted genes for completely sequenced eukaryotic organisms have no assigned function (Hanson et al., 2010).

- More than 5000 Arabidopsis genes are still hypothetical or unknown genes according the TAIR v10 annotation.

Our knowledge about gene candidates involved in the adaptation of plants to their environment remains partial (of potential interest for crop improvement).
Functional annotation procedures based on sequence similarities have reached their limitations.

One gene-one enzyme hypothesis is now considered as an oversimplification.

- Availability of thousands of transcriptomes: allow us to shift from a ‘gene by gene’ approach to more global approach through the ‘guilt by association’ concept.

- **Hypothesis:** Coexpressed genes have likely related biological functions (Eisen et al., 1998)
Coexpression analysis

Bioinformatic characterisation (GO, TFBS, PPI, TF-target, bibliome…)

Focus on specific genes for biological validation

---

**Drawbacks**

- Data are generally extracted from international repositories
- It leads to heterogeneous data in terms of acquisition and preprocessing.
- Coexpression generally done by analyzing gene pairs (Pearson correlation)
- It is a local point of view of a complex question.
Our Approach: Goals & methods

Goals:

I. Provide a global overview of the coexpression units of genes responding to a panel of stress stimuli in *Arabidopsis thaliana*.
II. Go beyond the coexpression to identify coregulated modules of functional partner genes.
III. Inference of function to orphan genes in well-characterized modules.

Methods: original features

- The specificities of the dataset and the biological theme: homogeneous and dedicated transcriptomic data.
- The method of clustering: model based method
- The integration of various resources to improve the functional inference.
Transcriptomic Dataset

CATdb: Complete Arabidopsis transcriptome database

- Large and homogeneous transcriptome resource generated by the CATMA platform of URGV and available in CATdb (http://urgv.evry.inra.fr/CATdb; Gagnot et al., NAR 2008).

- ~6000 original genes not present in the commonly used ATH1 Affymetrix DNA chip.

- All experiments dedicated to stresses were considered: 9 biotic and 9 abiotic stress categories

Differential expression analysis

17,264 genes have transcription ‘impacted’ (directly or not) by at least one stress experiment
CoExpression Analysis

**Gene Clustering:** Identification of co-expressed genes from the expression differences through a Model based clustering method for each stress category.

- **Matrix**
  \[
  \{ \text{genes x experiments} \} 
  \]
  By stress

- **Mathematical Criterion to select the cluster number (BIC)**

- **Classification rule based on conditional probabilities**

- **~700 Clusters of Coexpressed Genes**

<table>
<thead>
<tr>
<th>Stress category</th>
<th>Gene_nb</th>
<th>Clusters_nb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrogen</td>
<td>13 495</td>
<td>59</td>
</tr>
<tr>
<td>Temperature</td>
<td>11 365</td>
<td>34</td>
</tr>
<tr>
<td>Drought</td>
<td>8 143</td>
<td>34</td>
</tr>
<tr>
<td>Salt</td>
<td>5 729</td>
<td>30</td>
</tr>
<tr>
<td>Heavy metal</td>
<td>10 617</td>
<td>57</td>
</tr>
<tr>
<td>UV</td>
<td>7 894</td>
<td>37</td>
</tr>
<tr>
<td>Gamma</td>
<td>5 350</td>
<td>32</td>
</tr>
<tr>
<td>Oxydative stress</td>
<td>10 127</td>
<td>52</td>
</tr>
<tr>
<td>Nectrophic bacteria</td>
<td>11 220</td>
<td>50</td>
</tr>
<tr>
<td>Biotrophic bacteria</td>
<td>12 023</td>
<td>56</td>
</tr>
<tr>
<td>Fungi</td>
<td>9 773</td>
<td>51</td>
</tr>
<tr>
<td>Rhodococcus</td>
<td>1 900</td>
<td>13</td>
</tr>
<tr>
<td>Oomycete</td>
<td>5 508</td>
<td>31</td>
</tr>
<tr>
<td>Nematode</td>
<td>7 413</td>
<td>27</td>
</tr>
<tr>
<td>Stifenia</td>
<td>1 525</td>
<td>17</td>
</tr>
<tr>
<td>Virus</td>
<td>11 832</td>
<td>54</td>
</tr>
</tbody>
</table>
This project has been implemented as a new CATdb module: GEM2Net associated with a user-friendly Interface.

http://urgv.evry.inra.fr/GEM2NET
Functional inference by coregulation analysis

CoExpression clusters for each category of stress

Identification of Coregulated genes

Describe groups of functional partners

Annotation of orphan genes

Integration

CoRegulation Network
What is CoRegulation?

Group of genes which are coexpressed in several stress conditions: key players of stress response
CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions: key players of stress response

How to perform it?
CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions: key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories
What is CoRegulation?

Group of genes which are coexpressed in several stress conditions: key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories

Pairs conserved in at least n stresses

<table>
<thead>
<tr>
<th>n</th>
<th>Nbr pair of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5 533 013</td>
</tr>
<tr>
<td>3</td>
<td>423 771</td>
</tr>
<tr>
<td>4</td>
<td>68 875</td>
</tr>
<tr>
<td>5</td>
<td>19 113</td>
</tr>
<tr>
<td>6</td>
<td>6 987</td>
</tr>
<tr>
<td>7</td>
<td>3 366</td>
</tr>
<tr>
<td>8</td>
<td>1 679</td>
</tr>
<tr>
<td>9</td>
<td>786</td>
</tr>
<tr>
<td>10</td>
<td>324</td>
</tr>
<tr>
<td>11</td>
<td>171</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>39</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>
CoRegulation Analysis

What is CoRegulation?

Group of genes which are coexpressed in several stress conditions: key players of stress response

How to perform it?

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories

When coexpression becomes coregulation?

Pairs conserved in at least n stresses

<table>
<thead>
<tr>
<th>n</th>
<th>Nbr pair of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5 533 013</td>
</tr>
<tr>
<td>3</td>
<td>423 771</td>
</tr>
<tr>
<td>4</td>
<td>68 875</td>
</tr>
<tr>
<td>5</td>
<td>19 113</td>
</tr>
<tr>
<td>6</td>
<td>6 987</td>
</tr>
<tr>
<td>7</td>
<td>3 366</td>
</tr>
<tr>
<td>8</td>
<td>1 679</td>
</tr>
<tr>
<td>9</td>
<td>786</td>
</tr>
<tr>
<td>10</td>
<td>324</td>
</tr>
<tr>
<td>11</td>
<td>171</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>39</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>
CoRegulation Analysis

- We searched a threshold k at which the probability that a pair of genes is coexpressed in k stresses is significantly different from random.

- Permutation scheme:

  1. Do 1000 times
     a) Shuffle gene classification within each stress category.
     b) Occurrence calculation

  2. Error rate calculation: average of occurrence in random samples divided by the occurrence in our data.

<table>
<thead>
<tr>
<th>Number of stresses (n)</th>
<th>Random Network</th>
<th>Biological Network</th>
<th>Error_rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4+</td>
<td>1549</td>
<td>32313</td>
<td>4.79%</td>
</tr>
<tr>
<td>5+</td>
<td>12</td>
<td>13200</td>
<td>0.09%</td>
</tr>
<tr>
<td>6+</td>
<td>0</td>
<td>6216</td>
<td>0%</td>
</tr>
<tr>
<td>7+</td>
<td>0</td>
<td>3366</td>
<td>0%</td>
</tr>
</tbody>
</table>

Occurrence of pairs conserved in at least n stresses within our random and biological networks
Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

Legend
- Coregulated genes
- Orphan genes
- TF
CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF

Legend

- Coregulated genes
- Orphan genes
- TF
CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF

31 connected components

Legend

Coregulated genes
Orphan genes
TF
CoRegulation Networks

Pairs conserved in at least 7 stresses
867 genes, 3366 pairs

contains

178 Orphans
57 TF

31 connected components

Functional Modules ???

Legend

- Coregulated genes
- Orphan genes
- TF
Gene Ontology Enrichment Analysis

GO: Cellular Component Ontology

Specific and homogeneous modules
Cis-regulatory motifs Enrichment Analysis

Global CoRegulation Network Analysis (867 genes)

- 30 TFBS are found Over-represented by comparison with the whole genome present at most in 30% of promoters.

CoRegulation Network Analysis by component (9 largest components)

- 8 components are enriched in TFBS.
- 4 components are enriched with a pattern that is present in over 60% of their promoters
Cis-regulatory motifs Enrichment Analysis

Global CoRegulation Network Analysis
(867 genes)

- 30 TFBS are found Over-represented by comparison with the whole genome present at most in 30% of promoters.

CoRegulation Network Analysis by component
(9 largest components)

- 8 components are enriched in TFBS.
- 4 components are enriched with a pattern that is present in over 60% of their promoters

- Genes are under the control of the same common regulators
- Using the network topology is a good track for identifying modules
- Module = Functional partners
Example of an orphan identification within a module

98 genes, 1112 pairs, 1 orphan

First Neighbors of the orphan gene
Example of an orphan identification within a module

98 genes, 1112 pairs, 1 orphan

First Neighbors of the orphan gene

15 of the 17 neighbors are annotated as «Structural constituent of ribosome»
Example of an orphan identification within a module

98 genes, 1112 pairs, 1 orphan

First Neighbors of the orphan gene

15 of the 17 neighbors are annotated as « Structural constituent of ribosome »

This orphan gene most likely codes for a ribosomal protein
Identification of Coregulated genes → Describe groups of functional partners → Annotation of orphan genes

At Network «7+» stresses scale
Conclusion & Prospects

At Network «4+»
stresses scale

1. Identification of Coregulated genes
2. Describe groups of functional partners
3. Annotation of orphan genes
Conclusion & Prospects

At Network «4+» stresses scale

- Identification of Coregulated genes
- Describe groups of functional partners
- Annotation of orphan genes
Conclusion & Prospects

At Network « 4+ »
stresses scale

Identify Coregulated genes

Describe groups of functional partners

Annotation of orphan genes

- Refine the search of functional modules using methods of topological network analysis
- Integration with interactome data to improve the quality of function inference
- First step toward regulatory networks
Acknowledgements

Bioinformatics
ML.Martin-Magniette
S. Aubourg
V. Brunaud
G. Rigaill
J.-P. Tamby
C. Guichard
Z. Tariq

Transcriptome, mutants and stress
E. Delannoy
C. Lurin, S. Balzergue

Model-based clustering
G. Celeux (Orsay)
C. Maugis (Toulouse)

Funding:

GAP BV MIA
Thank you for your attention
Next steps: From clusters to gene networks

From co-regulation clusters to groups of putative functional partners
- Merging and extension of clusters (functions, profiles),
- Experimental step (validation): Transcriptomes of mutants (putative regulators and TF will be targeted)

**Gaussian Graphical Model**

Prediction of gene networks (conditional dependencies)

New candidate genes for crop improvement

Translational research

**CATdb**

Putative functional partners

**MIA**

**TILLing platform**

**URGV**

**Transcriptome platform**

**Giralud et al. (2009)**