Comparative genomics of gene families in relation with metabolic pathways for gene candidates highlighting

Delphine Larivière & David Couvin
Under the supervision of Dominique This, Jean-François Dufayard and Stéphanie Bocs
Summary

- Introduction to gene families
- Identification of evolutive events
- Metabolic pathways
- GenFam
- GenesPath
- Perspectives
- Conclusions
Gene families, synteny, and metabolic pathways
Introduction to Gene Families: Definition
Introduction to Gene Families: Evolution

- Speciations

- Duplications:
  - Genes
  - Chromosomal segments
  - Complete Genomes (WGD, polyploidy)
Introduction to Gene Families: Functional Evolution

- Type of Evolution
  - Neo-functionalization
  - Sub-functionalization
  - Pseudo-genes, Losses

Source: Pathway tool
Introduction to Gene Families: Hypothesis

- The function is more preserved between orthologs than between paralogs [Altenhoff et al., 2012]
- Need to consider the time of divergence between two sequences; indeed, close paralogues may have undergone less changes than ancient orthologs and therefore present more similar functions [Nehrt et al., 2011]

Knowledge of evolutive history of genes important for functional inference.
Analysis of evolutive event
Analysis of evolutive event: Synteny
Analysis of evolutive event: Synteny
Metabolic pathways

- In a metabolic pathway, the product of one enzyme acts as the substrate for the next. These enzymes often require dietary minerals, vitamins, and other cofactors to function.

- In our studies, the link between metabolic pathways and gene families is important to better understand the specificities of some genes, and their involvement in identified biological processes.

Example: Amphibolic Properties of the Citric Acid Cycle
GenFam: a tool dedicated to gene families study
GenFam: Analysis of gene Families

HMM Clustering
- Families
  - Alignment
  - Phylogénie
  - Réconciliation
- Multi species
  - Prediction of gene relationships

Blast
- Similarity
  - Filtration
  - Syntenic blocks
  - Syntenic depth
  - dS
- Pair of species

Functional evidences
- Visualization
GenFam: Analysis of gene Families

HMM Clustering
- Families
  - Alignment
  - Phylogénie
  - Réconciliation
- Prediction of gene relationships

Blast
- Similarity
  - Filtration
  - Syntenic blocks
  - Syntenic depth
  - dS
- Prediction of gene relationships

Manual Analysis
- Functional evidences
- Visualization

MULTIE SPECIES
IDEVEN
GenFam: Architecture
GenFam: Web Portal

- Creation of initial set of sequences
- Import through fasta or json format
  - GreenPhyl families
- Access to Chado Databases
  - Banana
  - Coffee
- Cart of sequences to manage the family

GenFam is available at: http://genfam.southgreen.fr/
GenFam: Homologs search
GenFam: Phylogeny
GenFam: Ideven for Synteny Integration

- 25 species
- Synteny identification through CoGe SynMap suite.
- dS Calculation (Haibao Tang)
  - Yang Nielson

<table>
<thead>
<tr>
<th>species</th>
<th>MAIZE</th>
<th>SCRM1</th>
<th>SETIT</th>
<th>ORYSJ</th>
<th>BRADJ</th>
<th>HORVU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCRM1</td>
<td>1/2</td>
<td>null</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
</tr>
<tr>
<td>SETIT</td>
<td>1/2</td>
<td>1/1</td>
<td>null</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
</tr>
<tr>
<td>ORYSJ</td>
<td>1/2</td>
<td>1/1</td>
<td>1/1</td>
<td>null</td>
<td>1/1</td>
<td>1/1</td>
</tr>
<tr>
<td>BRADJ</td>
<td>1/2</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>null</td>
<td>1/1</td>
</tr>
<tr>
<td>HORVU</td>
<td>1/2</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>null</td>
</tr>
</tbody>
</table>
## GenFam: Ideven for Synteny Integration

<table>
<thead>
<tr>
<th>#Nom gene1</th>
<th>Nom gene 2</th>
<th>Event</th>
<th>dS</th>
<th>Mean dS</th>
<th>Block size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G02280.1-PROTEIN_ARATH</td>
<td>BD01_PF61360.1_BRADI</td>
<td>ortholog</td>
<td>4.8013</td>
<td>3.7568285714285707</td>
<td>10</td>
</tr>
<tr>
<td>AT5G11110.1-PROTEIN_ARATH</td>
<td>BD03_PF06280.1_BRADI</td>
<td>ortholog</td>
<td>2.4897</td>
<td>2.3304153846153843</td>
<td>6</td>
</tr>
<tr>
<td>AT5G11110.1-PROTEIN_ARATH</td>
<td>BD03_PF20090.1_BRADI</td>
<td>ortholog</td>
<td>3.8082</td>
<td>3.1849600000000007</td>
<td>7</td>
</tr>
<tr>
<td>AT3G43190.1-PROTEIN_ARATH</td>
<td>GM09_PF06450.1_GLYMA</td>
<td>ortholog</td>
<td>1.4188</td>
<td>2.36896</td>
<td>7</td>
</tr>
<tr>
<td>GM13_PF08280.1_GLYMA</td>
<td>GM15_PF15670.1_GLYMA</td>
<td>WGD</td>
<td>0.6661</td>
<td>0.2959604095563141</td>
<td>146</td>
</tr>
</tbody>
</table>

---

*Note: The table above shows the results of the Ideven analysis for syntenic integration, including the gene names, orthologs, event types, dS values, mean dS values, and block sizes.*
**GenFam: Visualization**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Code</th>
<th>Function</th>
<th>Reviewed</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G10120.1_ARATH</td>
<td>SPSA</td>
<td>Probable sucrose phosphate synthase 4</td>
<td>reviewed</td>
<td>4/5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC 2.4.1.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sucrose phosphate synthase 4F</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGPS4F</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UDP-glucose-fructose-phosphate-glucosyltransferase</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Relationship</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G462613_T01_MAIZE</td>
<td>GRMZM2G049076_T01_MAIZE</td>
<td>WGD</td>
<td>0.8186</td>
</tr>
</tbody>
</table>

[Image of a Galaxy tool interface with a tree diagram and sequences highlighted.]
GenesPath: a tool using GenFam, and focused on metabolic pathways
GenesPath: the workflow

Uniprot

Uclust

GenFam + Phylogeny

Fasta files of protein sequences.
Query from keywords: Organisms (e.g. MAIZE, ORYSJ, ARATH, POPTR) + metabolic pathways

Clusters of homologous sequences in Fasta format

Galaxy Bricks
- MAFFT
- HMM build
- HMM search multi species
- HMM to fasta multi species
- Format fasta header
- MAFFT
- Gblocks
- Fasta2Phylip
- PhyML
- RAP-Green
GenesPath: web interface

Welcome To GenesPath

GenesPath (or gPTool) is an online tool allowing to highlight candidate genes in given plant species such as maize or sorghum. Unlike Gramene, EnsemblPlants and GreenPhyl, our approach brings more targeted responses to our queries focusing on specific proteins or metabolic pathways. Our approach provides more precision using GenFam (developed by Delphine Lariviére). Our tool also allows overlapping (a same gene can be found in different groups), genesPath uses robusts method to analyze and compare amino acid sequences (MAFFT, HHMER, Gblocks, ...) and build phylogeny (PhyML, RAP-Green)

GenesPath tool is linked to the Biomass For the Future (BFF) project. Further information are available here.
**GenesPath: query UniProt database**

Retrieve UniProt Entries corresponding to given plant species (e.g., maize or sorghum bicolor) and metabolic pathways (e.g., pathways involved in lignin biosynthesis).

Deletion of previously created files? [Delete]

Simple search
Enter a metabolic pathway or keyword: ___________________ Select a species: ________________ Submit

Multiple search
Enter a list of metabolic pathways (one per line): ___________________ Select a species: ________________ Submit

Search in function of EC number
EC: 1.11.1.7 Select a species: ________________ Submit

Search in function of Gene Ontology (GO)
Gene Ontology code: GO:0005576 Select a species: ________________ Submit
GenesPath: Launch workflow through the web

Do the Jobs!

1. Launch a job using a predefined group of sequences (single cluster)

Enter a fasta document containing at least 2 amino-acid sequences

Or upload it: Select file  / Upload Fasta

Select the parameters for your workflow, and indicate your email address:

GenFam's options
E-value (default value=0.000005): 
Databases or Species (download full list of available databases/species):

Gblocks options
Gblocks first parameter (eg. 51):  
Gblocks second parameter (eg. 70):  

RAP-Green's options
Gene threshold (eg. 0.95):  
Species threshold (eg. 10.0):  
Polymorphism threshold (eg. 0.00):  

Submit
GenesPath: phylogenetic tree
## GenesPath: CSV files representing orthologous genes

<table>
<thead>
<tr>
<th>Cluster name</th>
<th>Gene ID_Species</th>
<th>Gene Name</th>
<th>Locus Tag</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Orthologous group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>OS02G49230.1-PEP_ORYSJ</td>
<td>LOC_OS02G49230.1-PEP</td>
<td>OS02_PF33150.4_ORYSJ</td>
<td>ORYSJ02</td>
<td>30096306</td>
<td>30098580</td>
<td>1</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>SB10_PF10650.1_SORBI</td>
<td>SB10G010860.1</td>
<td>SB10_PF10650.1_SORBI</td>
<td>SORBI10</td>
<td>14421774</td>
<td>14424602</td>
<td>1</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>ZM05_PF39000.2_MAIZE</td>
<td>GRMZM2G075562_T01</td>
<td>ZM05_PF39000.2_MAIZE</td>
<td>MAIZE05</td>
<td>204171997</td>
<td>204174388</td>
<td>1</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>OS07G47140.1-PEP_ORYSJ</td>
<td>LOC_OS07G47140.1-PEP</td>
<td>OS07_PF28660.1_ORYSJ</td>
<td>ORYSJ07</td>
<td>28185202</td>
<td>28187762</td>
<td>2</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>ZM07_PF29680.1_MAIZE</td>
<td>GRMZM2G414423_T02</td>
<td>ZM07_PF29680.1_MAIZE</td>
<td>MAIZE07</td>
<td>171822613</td>
<td>171824015</td>
<td>2</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>SB02_PF34770.1_SORBI</td>
<td>SB02G042230.1</td>
<td>SB02_PF34770.1_SORBI</td>
<td>SORBI02</td>
<td>75932722</td>
<td>75934993</td>
<td>2</td>
</tr>
</tbody>
</table>
CONCLUSION AND PERSPECTIVES: GENFAM

Web system for manual analysis of gene families

Integration of syntenic analysis through IDEVEN for gene relationship prediction

Integration of heterogeneous data

Synthetic visualization in IntreeGreat

Use of Galaxy API to allow analysis through the GenFam website

Integration of dS based WGD identification

Integration of other types of functional evidences (TFBS)

Integration of syntenic and domains visualization to IntreeGreat
**Conclusion and perspectives: GenesPath**

Take into account gene families and metabolic pathways

Link with a specific project named “Biomass For the Future” (BFF)

GenFam and GenesPath are complementary tools

Launch the workflow of analysis without using Galaxy

Integration of phylogenetic analysis and statistics on genes (CSV) using IntreeGreat

GenesPath will be improved (it does not contain yet the IDEVEN tool)
Thank you for your attention!

genfam.southgreen.fr