GENE@HOME

GEne Network Expansion

Computational Biology project for GEne Networks Expansion on a Distributed Platform

@ **The 10th BOINC Workshop**
Budapest, 29 Sep 2014

Valter Cavecchia
National Research Council of Italy
CNR-IMEM, Trento Unit

TN-Grid BOINC platform

Who we are

Enrico Blanzieri

UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Claudio Moser

and the Gene Function Group

FONDAZIONE EDMUND MACH

140°

Valter Cavecchia

imem

Francesco Asnicar, Luca Masera Paolo Morettin, Nadir Sella, Thomas Tollio and all the students of the *Laboratory of Biological Data Mining* class, UniTN, 2013-2014

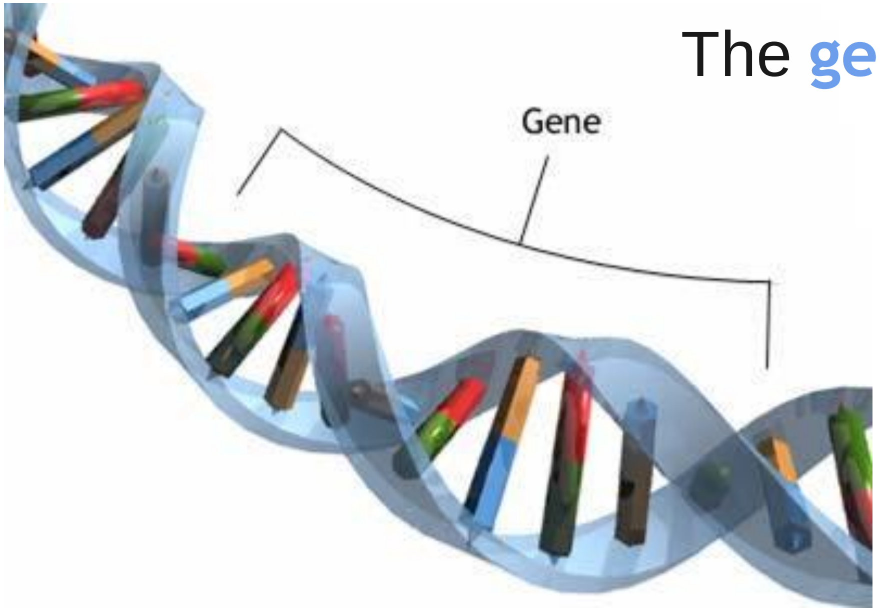
Testing hardware partially provided by CNR-ISTC-LOA, Trento

Biological background

A gene is a piece of DNA which contains the information to create a specific protein

The genome is the whole set of genes of a specific organism

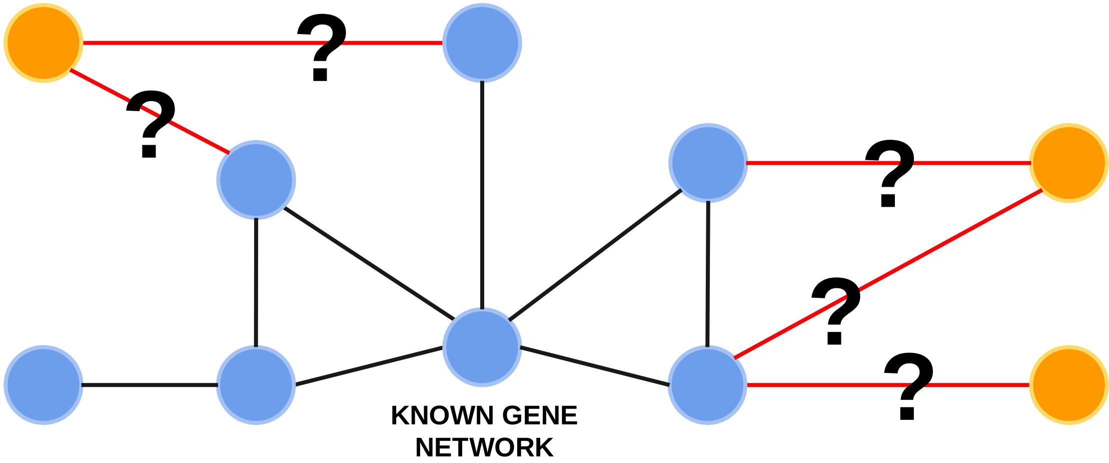
A subset of functionally interacting genes form a Local Gene Network (LGN)



GENE@HOME
GEne Network Expansion

Challenge

We want to discover **new relations** between genes (expansion)

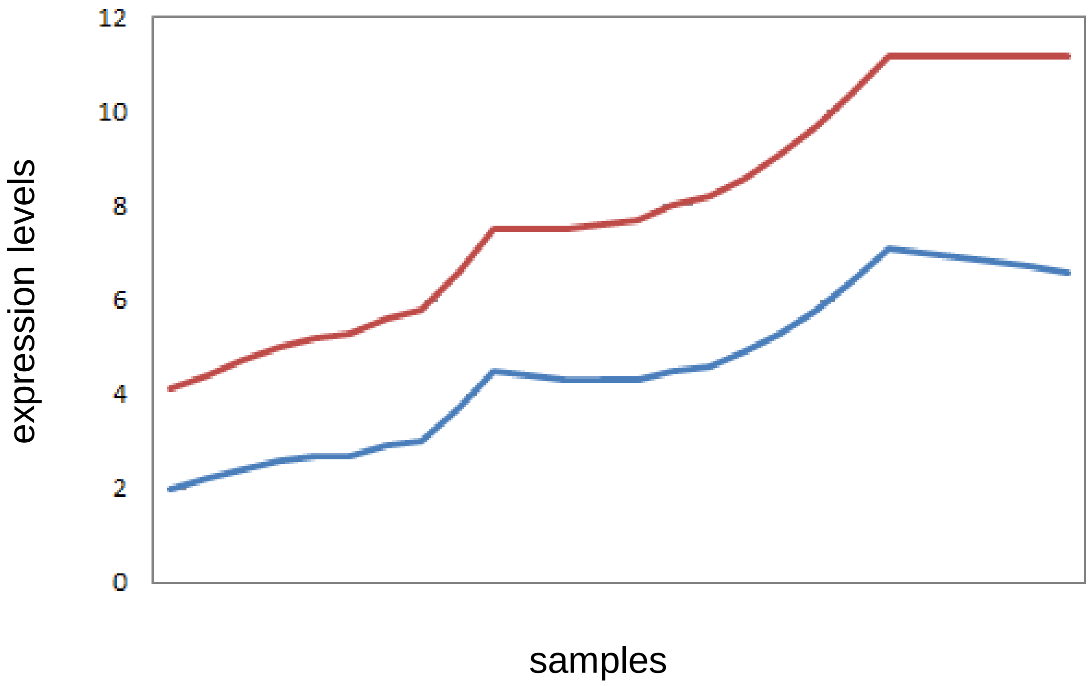


Genes on the same local gene network are **correlated**

Method

We compare the **expression levels** of two different genes

Relations between genes become **correlations** when their expression levels have a similar trend

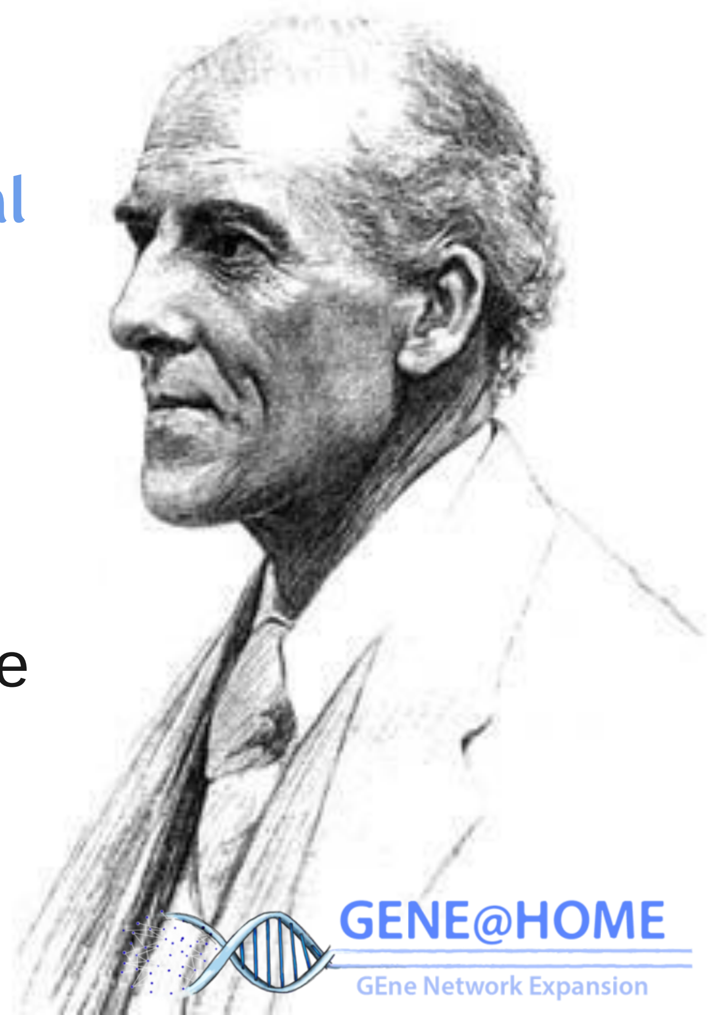


Method

We use the **PC-algorithm** to find **causal relationships** among genes, exploiting their expression levels in different samples

Correlations (linear) between genes are computed using **Pearson coefficient**

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$



GENE@HOME

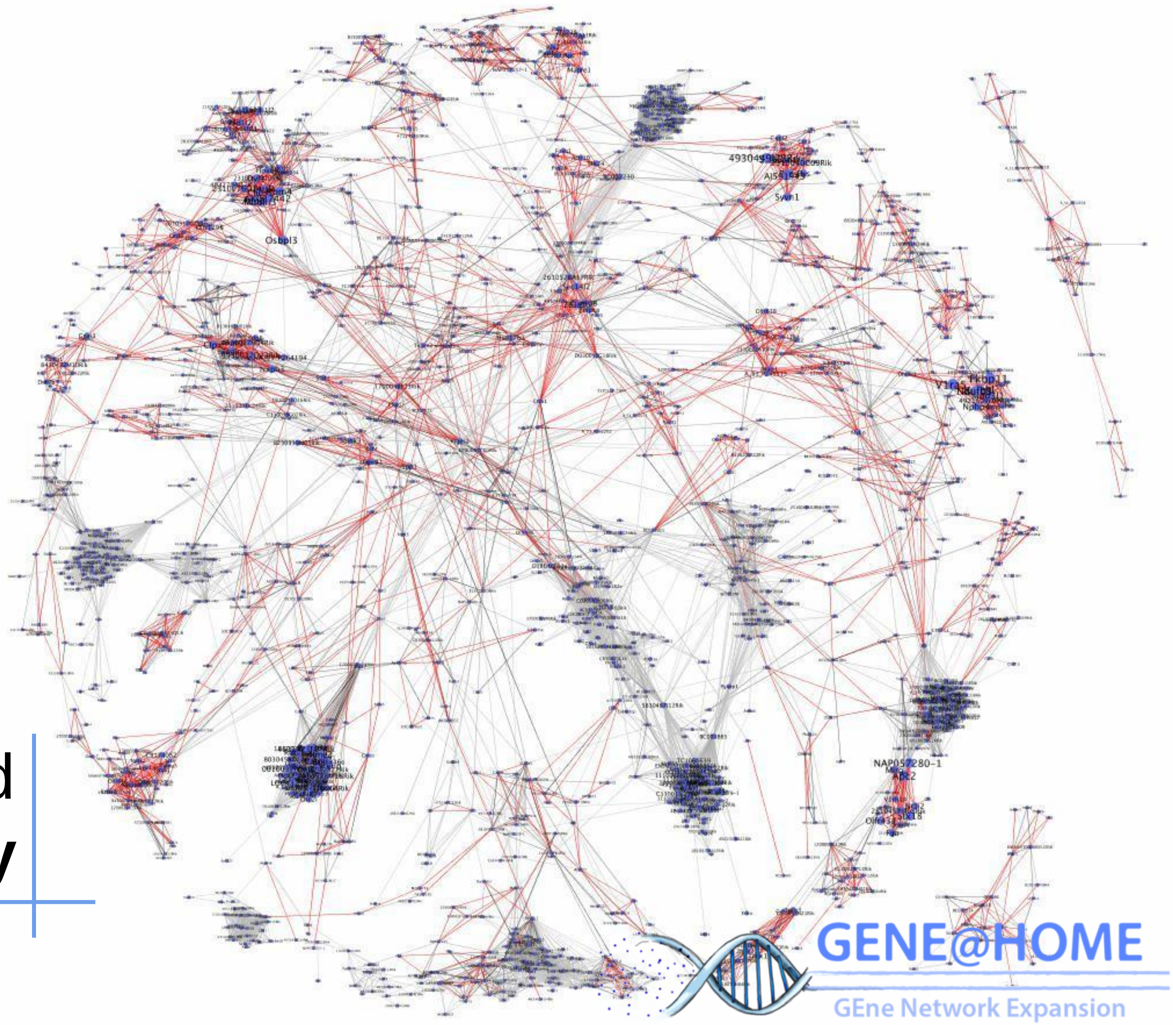
GEne Network Expansion

Problem

Genomes and gene networks are **huge**

We want to expand **many** local gene networks of **several** organisms

This work is **hard** and computationally **heavy**



GENE@HOME

GEne Network Expansion

Model

Study case

Arabidopsis thaliana

the model plant

~23.000 genes

~264.500.000 possible relations

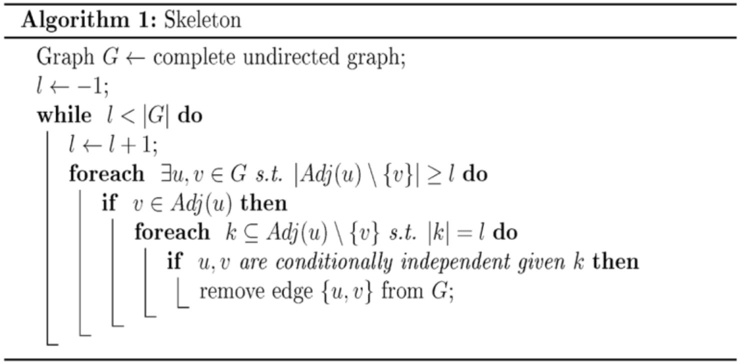
GENE@HOME

GEne Network Expansion

Implementation

1

Running the PC-algorithm on the whole genome is heavy. So we use **PC-IM** to iteratively run it on genome portions



```
Algorithm 1: Skeleton
Graph $G \leftarrow$ complete undirected graph;
$l \leftarrow -1$;
while $l < |G|$ do
    $l \leftarrow l + 1$;
    foreach $\exists u, v \in G$ s.t. $|Adj(u) \setminus \{v\}| \geq l$ do
        if $v \in Adj(u)$ then
            foreach $k \subseteq Adj(u) \setminus \{v\}$ s.t. $|k| = l$ do
                if $u, v$ are conditionally independent given $k$ then
                    remove edge $\{u, v\}$ from $G$;
```

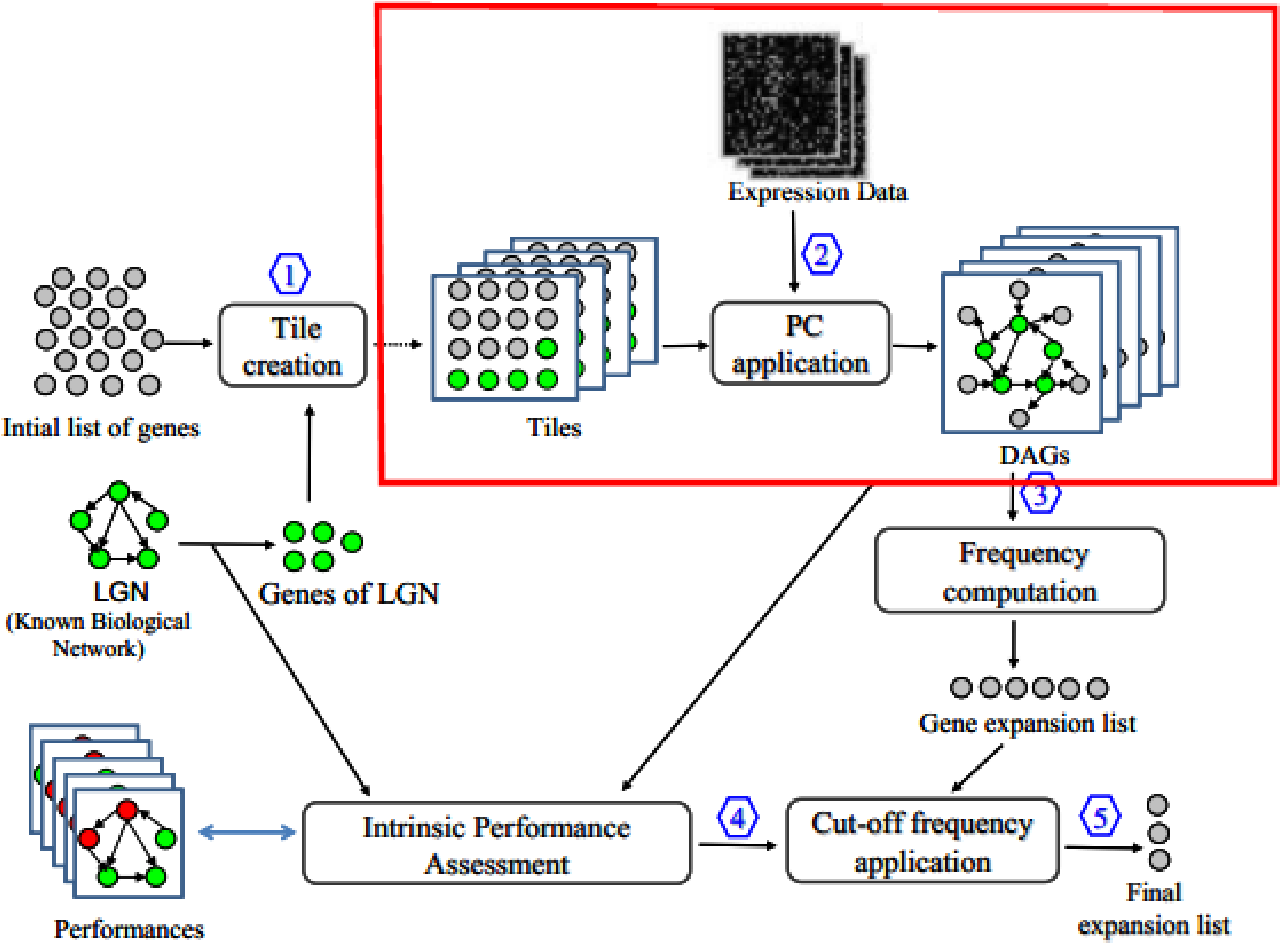
We implemented an efficient version of the PC-algorithm, named **PC++**

2

GENE@HOME

GEne Network Expansion

PC-IM


Expression Data
1
2
3
4
5
Tile creation
Tiles
PC application
DAGs
Intial list of genes
LGN
(Known Biological Network)
Genes of LGN
Frequency computation
Gene expansion list
Performances
Intrinsic Performance Assessment
Cut-off frequency application
Final expansion list

Implementation

We need a lot of computational power

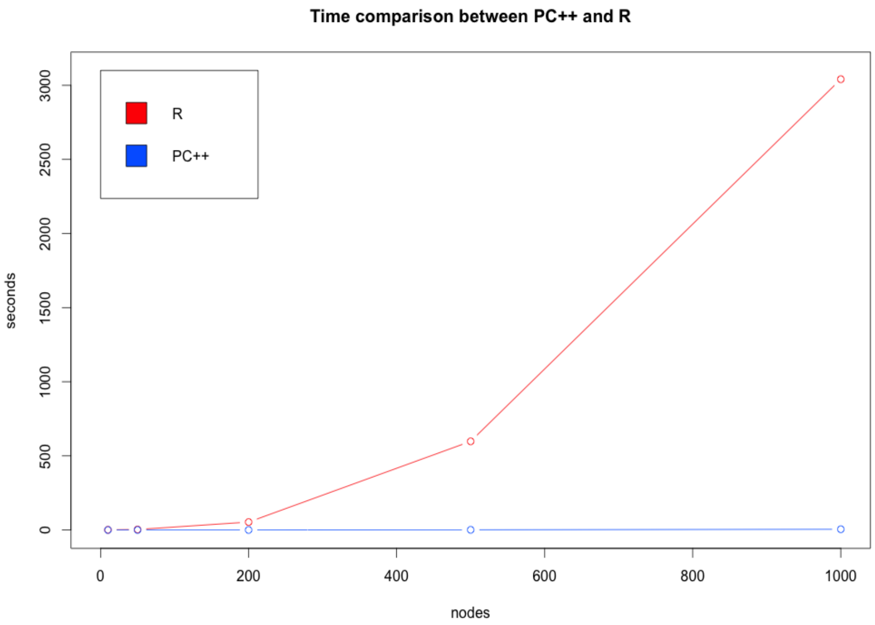
3 We use **BOINC**, an open source framework for Volunteer Grid Computing.

Thanks to the help of volunteers, we reached the computational power of a supercomputer

Implementation

R ⇨ C++ (Dynamic Programming, Adjacency Matrix)

$$\rho_{i,j|k} = \frac{\rho_{i,j|k\setminus h} - \rho_{i,h|k\setminus h}\rho_{j,h|k\setminus h}}{\sqrt{(1-\rho^2_{i,h|k\setminus h})(1-\rho^2_{j,h|k\setminus h})}}$$ $O(3^l)$



Algorithm 2: Correlation

```
function Dynamic correlation (int $l$, matrix $\rho$)
    $dim \leftarrow l + 2$;
    for $k = 1$ to $l$ do
        for $i = 0$ to $l - k$ do
            for $j = i + 1$ to $dim - k$ do
                $\rho[i][j] = \rho[j][i] = \frac{\rho[i][j] - \rho[i][dim-k] * \rho[j][dim-k]}{\sqrt{(1-\rho^2[i][dim-k]) * (1-\rho^2[j][dim-k],2)}}$;
    return $\rho[0][1]$;
```

$O(l^3)$

Boinc integration

BOINC API

Checkpoints

Running time estimates

- 20m-20h runtime

Memory, network and storage

- Implementation focused to minimize RAM usage and bandwidth
- gzip file transfer, *sticky* files

Multi-platform porting issues

- erf() function etc... (MS VisualC++ vs g++)

Supported Operating Systems

- Windows (x32/x64) from XP
- Mac OS X (CPU Intel, x64) version >= 10.5
- GNU/Linux (x32/x64) from kernel 3.x

Recommended Boinc client version: 7.0+

Boinc integration

Validation

- Simple bitwise (gzip version) validator
- Simple redundancy with `min_quorum = 2`

Work Generator

- Python scripts (may be improved)

Scheduler

- Standard (was using homogeneous redundancy)

Approach

- Alpha stage (internal)
- Beta stage (with invitation code, per request)

Issues (to-do list)

- Upgrade server (now virtual, with limited resources)
- Automation of post-processing phase
- Web (easy) access to job generation
- GPGPU version? (PC*)

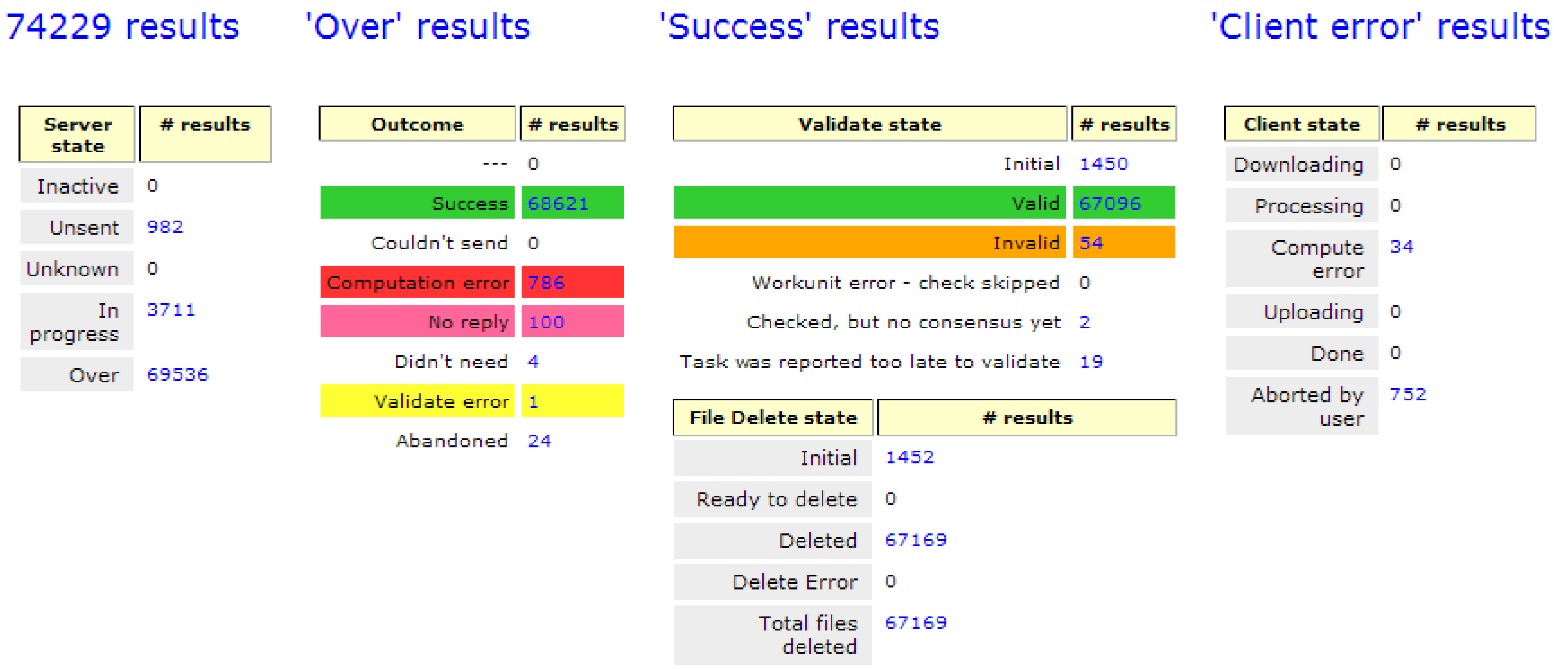
URL

- http://gene.disi.unitn.it/test/index.php

Boinc results

Users

- ~150 with ~550 hosts



74229 results

Server state	# results
Inactive	0
Unsent	982
Unknown	0
In progress	3711
Over	69536

'Over' results

Outcome	# results
---	0
Success	68621
Couldn't send	0
Computation error	786
No reply	100
Didn't need	4
Validate error	1
Abandoned	24

'Success' results

Validate state	# results
Initial	1450
Valid	67096
Invalid	54
Workunit error - check skipped	0
Checked, but no consensus yet	2
Task was reported too late to validate	19

File Delete state	# results
Initial	1452
Ready to delete	0
Deleted	67169
Delete Error	0
Total files deleted	67169

'Client error' results

Client state	# results
Downloading	0
Processing	0
Compute error	34
Uploading	0
Done	0
Aborted by user	752

Experiments

Organism

- *Arabidopsis Thaliana* Gene Expression Data
- 393 hybridization experiments

Local Gene Network

- Flower Organ Specification Gene Regulatory Network (FOS)
- 15 genes linked by 54 causal relationships

Experiments

- Precision
- Performance benchmark against competitors
- Sensitivity to algorithm parameters:
  - t - tile size
  - i - iterations
  - α - significance level
- Post-processing - (k) genes to be considered in the output list

Example (experiment 1, precision)

- ($\alpha = 0.05$) (t = 50; 100; 250; 500; 750; 1000; 1250; 1500; 1750; 2000) (i = 20; 50; 100; 250; 500; 1000; 1500; 2000)

Example (experiment 6, sensitivity)

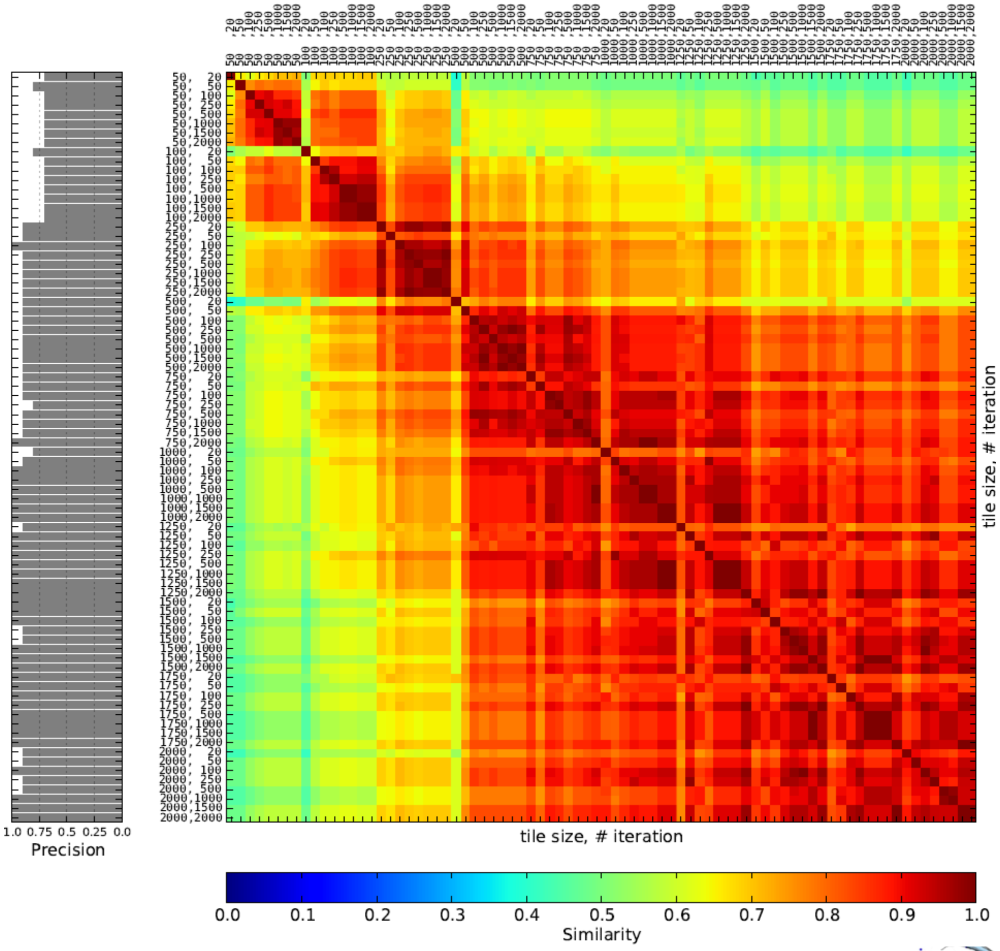
- “Leave one out” (14 genes out of 15)
- ($\alpha = 0.01,0.05$) (t = 1000, 2000, 3000, 4000) (I = 100, 2000).

Scientific results

Experiment 1

α = 0.05

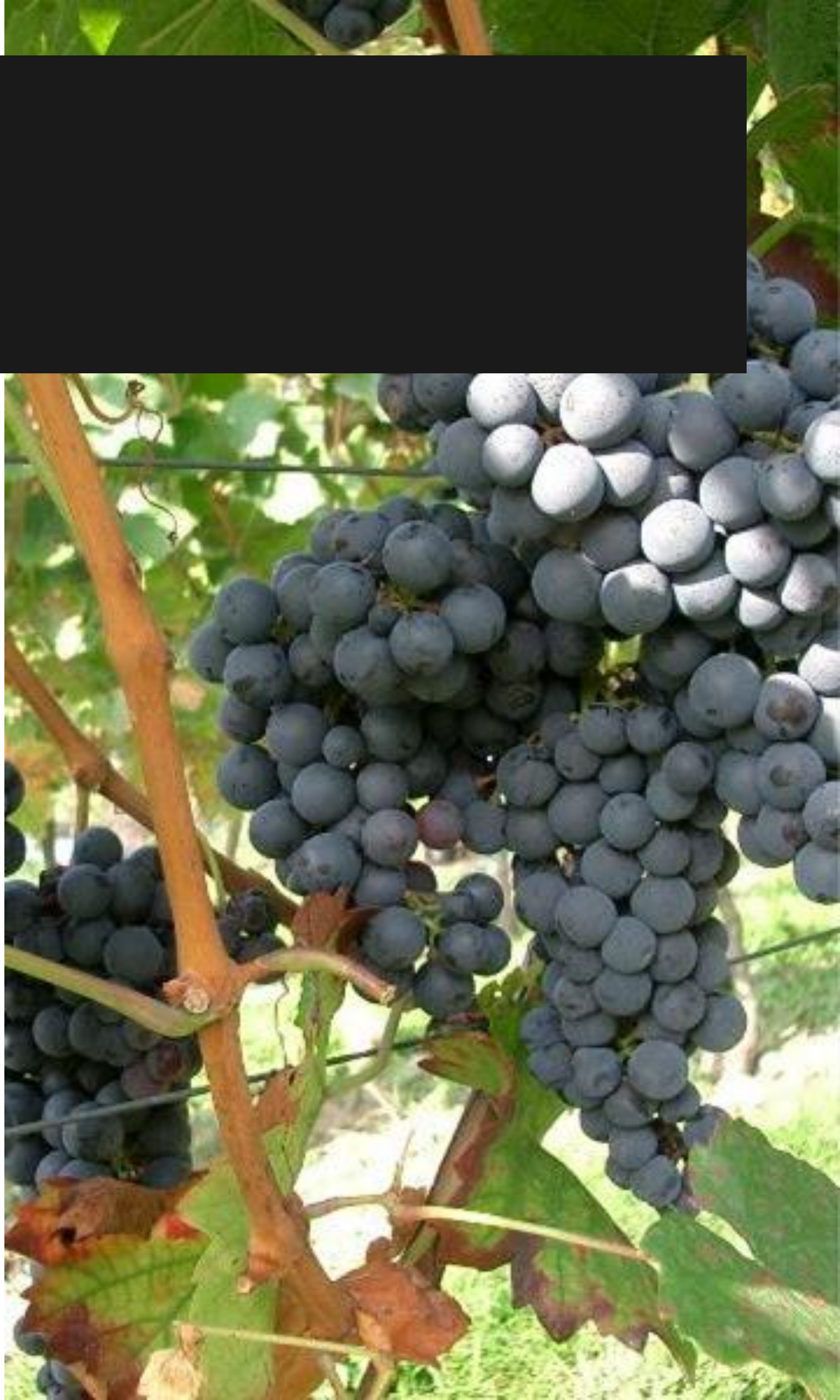
top 10 results (k)



Future work

Other organisms, other LGNs, focus on ‘regional’ agriculture

- *Escherichia coli* (bacteria)
- *Saccharomyces cerevisiae* (yeast)
- *Vitis vinifera* (grapevine)
- *Malus domestica* (apple)
- *Homo sapiens* (human)
- *Drosophila suzuki* (fruitfly)



Questions are welcome

GENE@HOME
GEne Network Expansion