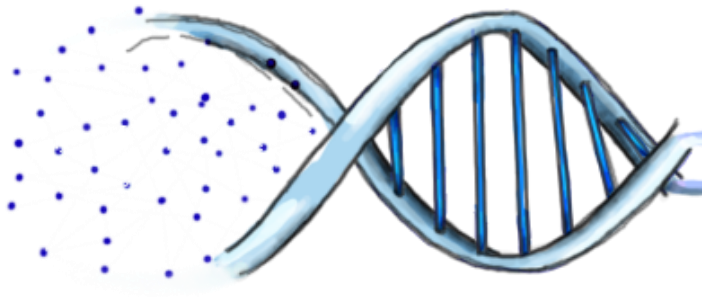


GENE@HOME

GEne Network Expansion

TN-Grid and gene@home project: Volunteer Computing for Bioinformatics

@ **BOINC:FAST Conference**
Petrozavodsk, 15th Sep 2015

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

TN-Grid BOINC platform

Who we are

Enrico Blanzieri

Claudio Moser

and the Gene Function Group

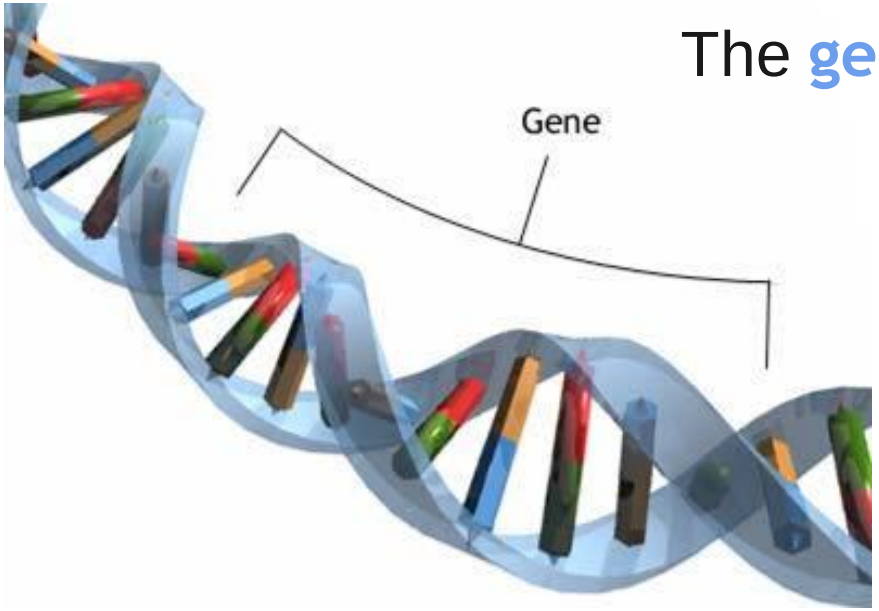
Valter Cavecchia

Francesco Asnicar, Luca Masera
Paolo Morettin, Nadir Sella, Thomas Tollio,
Stanislau Semeiuta and all the students of the *Laboratory of Biological Data Mining* class, UniTN, 2013-2015

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

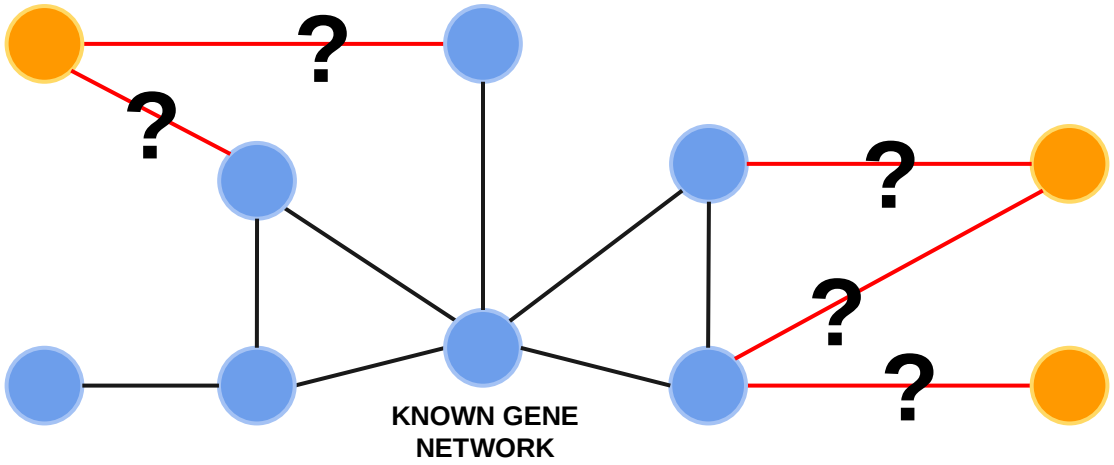


The **genome** is the whole set of genes of a specific organism

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

Challenge

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



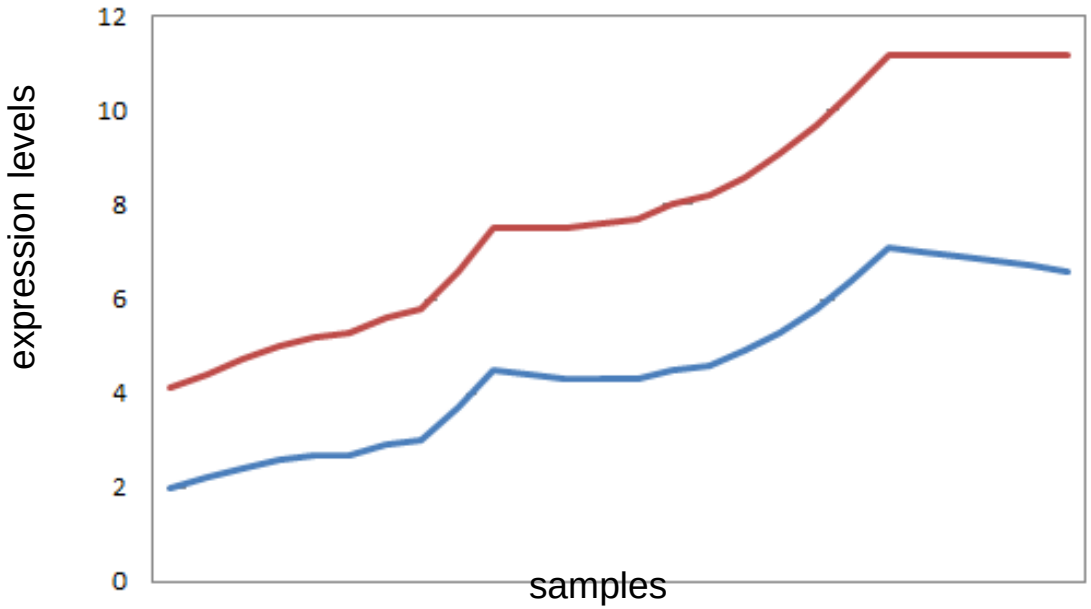
Genes on the same local gene network are **correlated**

GENE@HOME
GEne Network Expansion

Method

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

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



GENE@HOME
GEne Network Expansion

Method

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

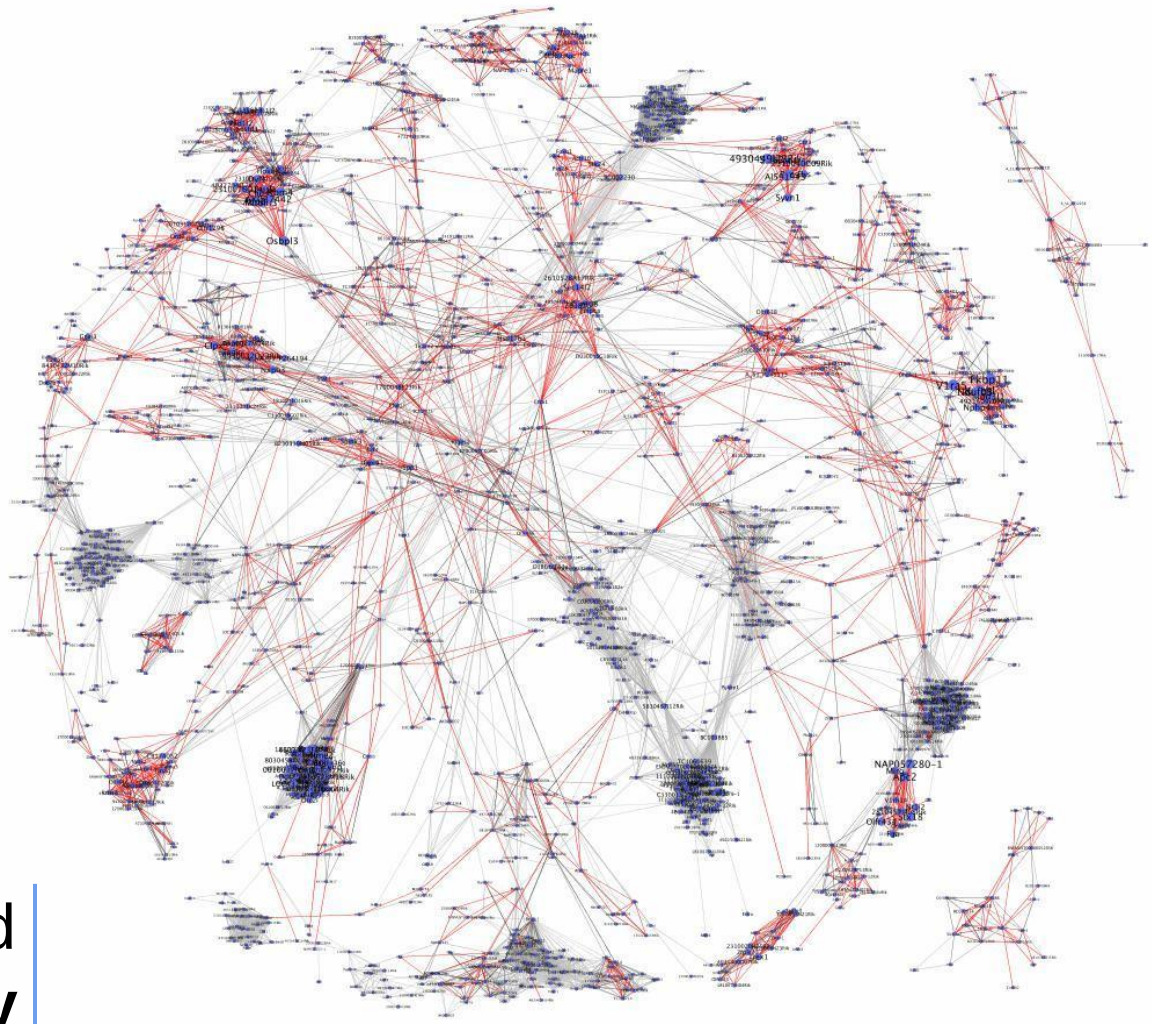
Correlations (linear) between genes are computed using **Pearson coefficient**

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

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

Arabidopsis thaliana
the model plant
~23.000 genes
~264.500.000 possible relations

Implementation

1

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

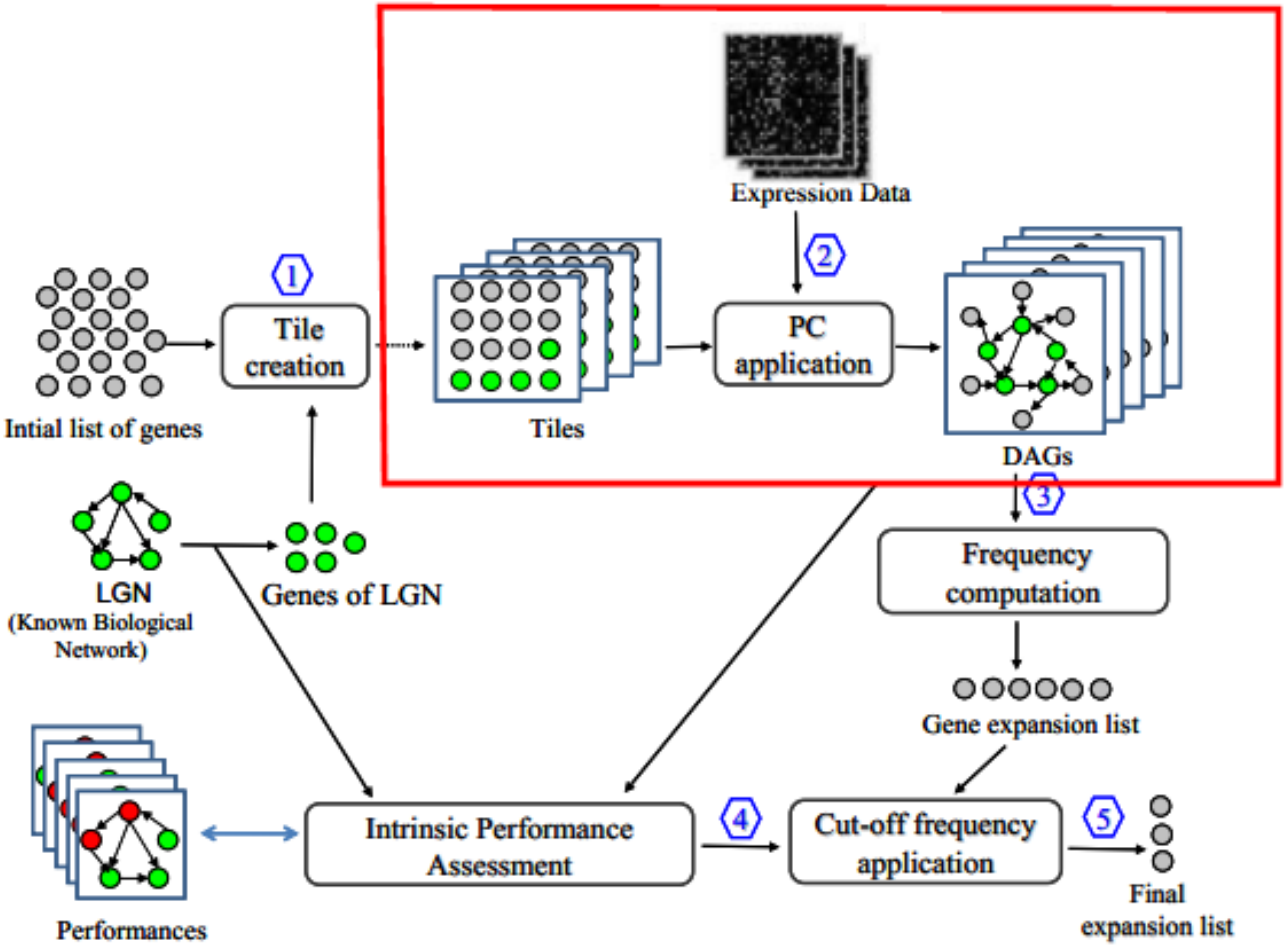
Algorithm 1: Skeleton

```
Graph G ← complete undirected graph;
l ← −1;
while l < |G| do
    l ← l + 1;
    foreach ∃u, v ∈ G s.t. |Adj(u) \ {v}| ≥ l do
        if v ∈ Adj(u) then
            foreach k ⊆ Adj(u) \ {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

PC-IM



GENE@HOME
GEne Network Expansion

Implementation

We need a lot of computational power

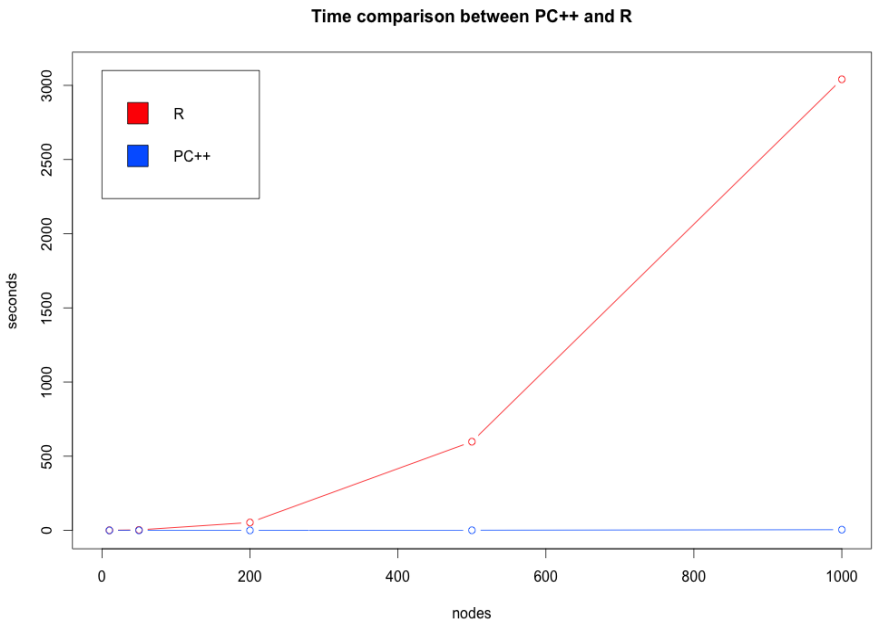
3

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)



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

Algorithm 2: Correlation

```
function Dynamic correlation (int l, matrix ρ)
    dim ← l + 2;
    for k = 1 to l do
        for i = 0 to l − k do
            for j = i + 1 to dim − k do
                ρ[i][j] = ρ[j][i] = (ρ[i][j] − ρ[i][dim−k] ∗ ρ[j][dim−k]) / sqrt((1−ρ²[i][dim−k]) ∗ (1−ρ²[j][dim−k]), 2);
    return ρ[0][1];
```

$O(l^3)$

GENE@HOME
GEne Network Expansion

Boinc integration

The 1st project hosted by the TN-Grid platform
The only Boinc active project in Italy (as of today)

- 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 and platforms
  - 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
- Fix validation issues (although rare)
- Web (more easier) access to job generation
- GPGPU version? (PC*)
- Send timed-out workunits to reliable hosts (Accelerating retries**)**

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

Boinc add-ons

gene@home PC-IM status

Monitoring tool written in php
Similar tool for whole history

Summary (including last 45 days)				
Total	Executed	In execution	Errors	Queued
122	54	8	0	60

Show 25 entries Search: Copy Print CSV

PC-IM now in execution (or queued)									
id	org	lgn	Last update	alpha	t-size	iter	pri	nWUs	Wait
4235	Ec	ecl_b0730-mngR	2015-09-01 01:58:59	0.050	200	2000	4	840	move
4236	Ec	ecl_b0817-mntR	2015-09-01 21:35:43	0.050	200	2000	4	840	2
4238	Ec	ecl_b0995-torR	2015-09-04 08:29:56	0.050	200	2000	4	880	3
4239	Ec	ecl_b1014-putA	2015-09-05 09:56:40	0.050	200	2000	4	840	15
4240	Ec	ecl_b1299-puuR	2015-09-06 12:14:54	0.050	200	2000	4	880	46
4241	Ec	ecl_b1303-pspF	2015-09-07 11:42:28	0.050	200	2000	4	880	173
4242	Ec	ecl_b1323-tyrR	2015-09-08 15:43:09	0.050	200	2000	4	880	589
4243	Ec	ecl_b1328-pgrR	2015-09-10 00:50:32	0.050	200	2000	4	840	840
4244	Ec	ecl_b1399-paaX	2015-07-31 18:43:40	0.050	200	2000	4	?	start
4245	Ec	ecl_b1450-mcbR	2015-07-31 18:43:40	0.050	200	2000	4	?	start
4246	Ec	ecl_b1499-ydeO	2015-07-31 18:43:40	0.050	200	2000	4	?	start
4247	Ec	ecl_b1512-lsrR	2015-07-31 18:43:40	0.050	200	2000	4	?	start
4248	Ec	ecl_b1530-marR	2015-07-31 18:43:40	0.050	200	2000	4	?	start
4249	Ec	ecl_b1564-relB	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4250	Ec	ecl_b1594-mlc	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4251	Ec	ecl_b1618-uidR	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4252	Ec	ecl_b1620-malI	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4253	Ec	ecl_b1642-slyA	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4254	Ec	ecl_b1649-nemR	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4255	Ec	ecl_b1827-kdgR	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4256	Ec	ecl_b1916-sdiA	2015-07-31 18:43:41	0.050	200	2000	4	?	start
4257	Ec	ecl_b2017-yefM	2015-08-08 11:53:22	0.050	200	2000	4	?	start
4258	Ec	ecl_b2105-rcnR	2015-08-08 11:53:22	0.050	200	2000	4	?	start
4259	Ec	ecl_b2125-yehT	2015-08-08 11:53:22	0.050	200	2000	4	?	start
4260	Ec	ecl_b2127-mlrA	2015-08-08 11:53:22	0.050	200	2000	4	?	start

Showing 1 to 25 of 68 entries Previous 1 2 3 Next

PC-IMs waiting to be moved: 1 (840 results)

Loading time 0.03 seconds, 5213 files (2015-09-10 15:28:31 UTC)

GENE@HOME
GEne Network Expansion

Boinc add-ons

gene@home hosts

Monitoring tool written in php

Checks timing and credit issues, will be expanded for *reliability* testing

Intel Core i5-2400 CPU@3.10GHz

Host #26 (nico) 3318.26 Mflops, 19768.18 Miops

Show 25 entries Search: Copy Print CSV

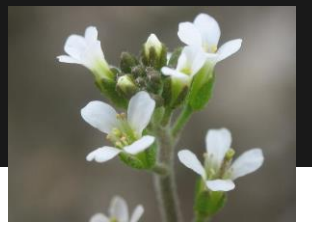
Results

id	org	lgn	alpha	tile-size	iterations	count	avg time	avg credit
4213	Ec	ecl_b3743-asnC	0.050	200	2000	2	5,224.96	58.23
4225	Ec	ecl_b4498-gatR	0.050	200	2000	1	5,080.38	65.17
4226	Ec	ecl_b0020-nhaR	0.050	200	2000	1	5,155.99	60.52
4229	Ec	ecl_b0305-rclR	0.050	200	2000	15	5,274.39	61.02
4231	Ec	ecl_b0345-lacI	0.050	200	2000	2	5,340.10	59.45
4232	Ec	ecl_b0346-mhpR	0.050	200	2000	14	5,353.92	58.17
4234	Ec	ecl_b0694-kdpE	0.050	200	2000	1	5,453.96	62.53
4235	Ec	ecl_b0730-mngR	0.050	200	2000	19	5,506.75	63.04
4236	Ec	ecl_b0817-mntR	0.050	200	2000	28	5,260.31	60.32
4237	Ec	ecl_b0846-rcdA	0.050	200	2000	19	5,109.58	58.62
4238	Ec	ecl_b0995-torR	0.050	200	2000	19	4,901.12	54.63
4239	Ec	ecl_b1014-putA	0.050	200	2000	20	5,316.22	59.55
4240	Ec	ecl_b1299-puuR	0.050	200	2000	28	5,011.58	55.66
4241	Ec	ecl_b1303-pspF	0.050	200	2000	26	5,195.52	56.87
4242	Ec	ecl_b1323-tyrR	0.050	200	2000	3	5,236.13	56.88

Showing 1 to 15 of 15 entries Previous 1 Next

Loading time 0.03 seconds (2015-09-10 15:45:56 UTC)

A. Thaliana



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)

- (α = 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)
- (α = 0.01,0.05) (t = 1000, 2000, 3000, 4000) (I = 100, 2000).

A. Thaliana

example data (expression levels)

```
probe,          Col-0-1, Col-0-2, Col-3-1, Col-3-2, Col-5-1, Col-5-2, Col-7-1, Col-7-2, ..
244901_at,   5.5707,  6.03712, 5.58236, 5.70144, 5.66002, 5.69186, 6.17956, 5.68236, ..
244902_at,   6.90945, 7.41808, 6.97188, 6.92903, 6.51446, 6.60624, 7.41142, 6.89409, ..
244903_at,   9.82084, 9.53556, 9.85939, 10.4963, 9.8111,  9.8772,  10.3531, 9.82263, ..
244904_at,   7.32661, 6.81494, 7.50378, 8.26042, 7.5489,  7.50737, 7.49373, 7.1771,  ..
244905_at,   4.78551, 5.29083, 4.76787, 5.12271, 4.92505, 5.11024, 5.1082,  4.68376, ..
244906_at,   6.68507, 6.91288, 6.62527, 6.91232, 7.17307, 7.22668, 7.14353, 6.76882, ..
244907_at,   4.3062,  4.46902, 4.01425, 4.86695, 4.29824, 4.12153, 4.80721, 4.18643, ..
244908_at,   4.53637, 4.77786, 4.34215, 4.98202, 4.59647, 4.61787, 4.89136, 4.35368, ..
244909_at,   4.68011, 5.02001, 4.69071, 5.2242,  4.64948, 4.58321, 5.13236, 4.37841, ..
244910_s_at, 4.86281, 5.12034, 4.6854,  5.40374, 4.68004, 4.61862, 5.6629,  4.76772, ..
244911_at,   4.08497, 3.97419, 3.78696, 4.3293,  3.63383, 3.61622, 4.50863, 3.92685, ..
244912_at,   10.8078, 11.0856, 10.8031, 10.9747, 11.0168, 11.2904, 11.0358, 10.8949, ..
244913_at,   6.39199, 6.50671, 6.51775, 6.88375, 6.64761, 6.59693, 6.42699, 6.06315, ..
..
..
```

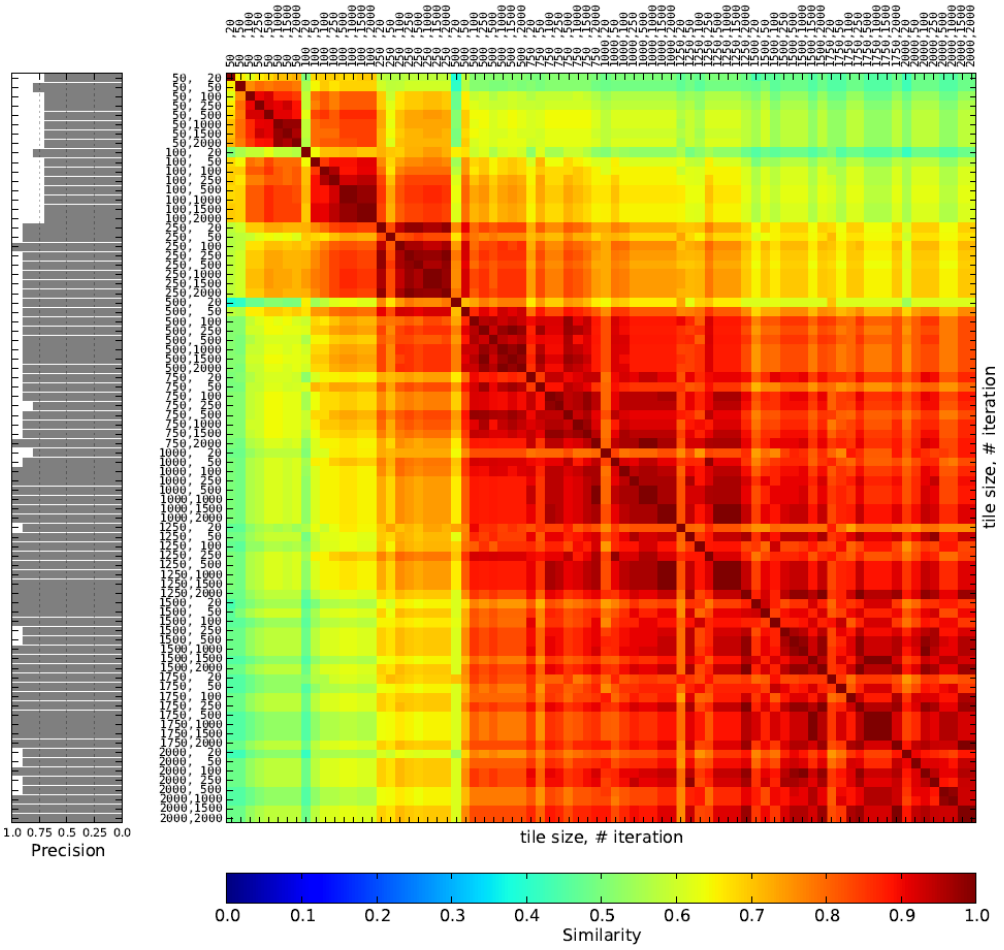
22810 probes

393 hybridizations

GENE@HOME
GEne Network Expansion

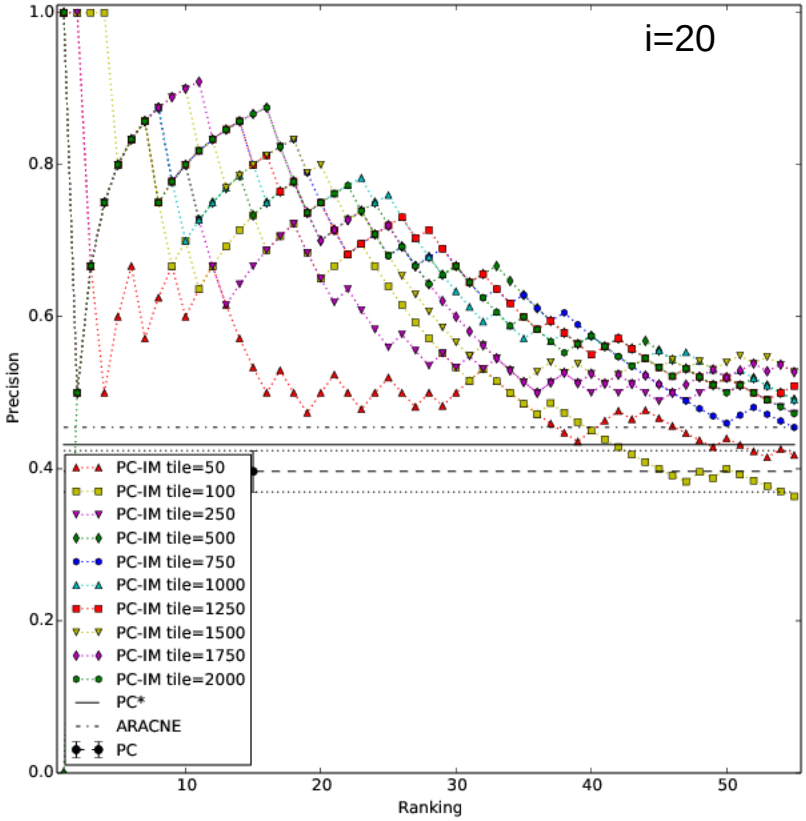
A. Thaliana: results

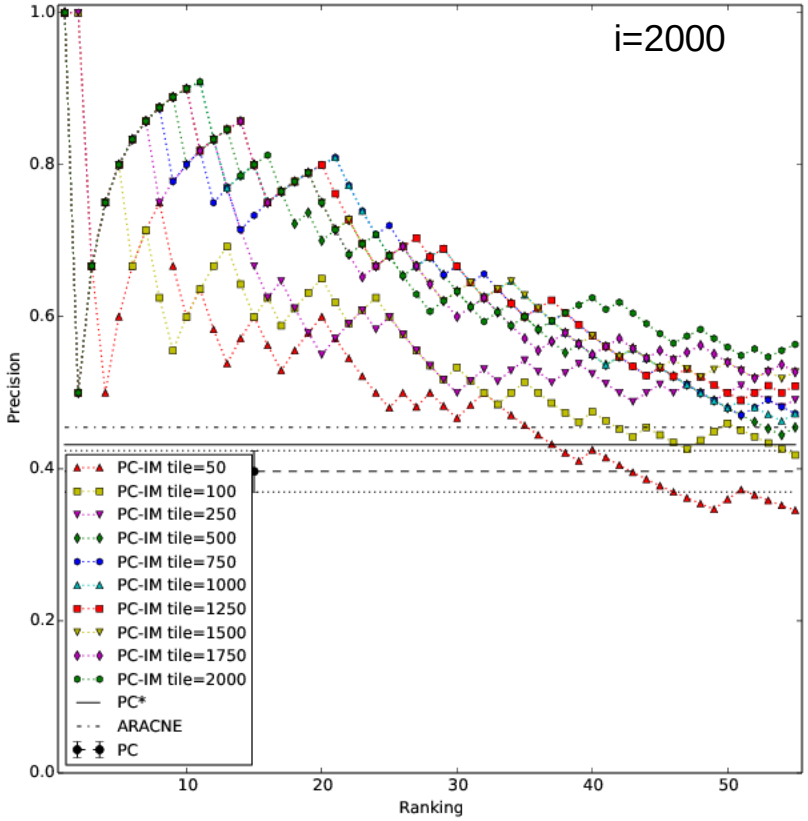
Experiment 1
$\alpha = 0.05$
top 10 results (k)



GENE@HOME
GEne Network Expansion

A. Thaliana: results





Precision comparison of A. thaliana data on FOS network. PC (average and variance), PC*, ARACNE, and PC-IMs with different levels of tile dimension with fixed α=0.05, (first 55 genes)

GENE@HOME
GEne Network Expansion

E. coli

1. Get expression data (from COLOMBOS):
   ec_v3f_mgn.csv ~100Mb (*sticky* file)
   4065 probes, max 2486 hybridizations (after *filtering*)

2. Parameter tuning:
   Selected LGNs (gene sets), "*empty*" and "*one gene*"
   (α = 0.01,0.05) (t = 100,200) (i =80,100,500,1000,1500,2000)

3. Experiments:
   144 LGNs (gene sets) (α = 0.05) (t = 200) (i = 2000)
   50 PC/workunit -> ~860 workunits every PC-IM
   ~1PC-IM per day @ ~520 GFlops

b0020-nhaR	b0566-envY	b1399-paaX	b2017-yefM	b2664-csiR	b3075-ebgR	b3574-yiaJ	b4063-soxR
b0034-caiF	b0571-cusR	b1450-mcbR	b2079-baeR	b2669-stpA	b3094-exuR	b3601-mtlR	b4089-alsR
b0069-sgrR	b0620-dpiA	b1499-ydeO	b2105-rcnR	b2684-mprA	b3118-tdcA	b3604-lldR	b4113-basR
b0162-cdaR	b0694-kdpE	b1508-hipB	b2125-yehT	b2697-alaS	b3119-tdcR	b3669-uhpA	b4116-adiY
b0226-dinJ	b0730-mngR	b1512-lsrR	b2127-mlrA	b2706-gutM	b3131-agaR	b3702-dnaA	b4118-melR
b0294-matA	b0817-mntR	b1530-marR	b2151-galS	b2707-gutR	b3226-nanR	b3743-asnC	b4124-dcuR
b0305-rclR	b0840-deoR	b1564-relB	b2163-yeiL	b2709-norR	b3255-accB	b3753-rbsR	b4133-cadC
b0313-betI	b0846-rcdA	b1570-dicA	b2213-ada	b2714-ascG	b3264-envR	b3773-ilvY	b4187-aidB
b0330-prpR	b0995-torR	b1594-mlc	b2217-rcsB	b2783-mazE	b3292-zntR	b3828-metR	b4191-ulaR
b0338-cynR	b1014-putA	b1608-rstA	b2220-atoC	b2805-fucR	b3418-malT	b3905-rhaS	b4241-treR
b0345-lacI	b1111-comR	b1618-uidR	b2289-lrhA	b2808-gcvA	b3422-rtcR	b3906-rhaR	b4260-pepA
b0346-mhpR	b1162-bluR	b1620-malI	b2364-dsdC	b2839-lysR	b3423-glpR	b3934-cytR	b4264-idnR
b0413-nrdR	b1201-dhaR	b1642-slyA	b2381-ypdB	b2980-glcC	b3438-gntR	b3938-metJ	b4324-uxuR
b0435-bolA	b1299-puuR	b1649-nemR	b2405-xapR	b3010-yqhC	b3481-nikR	b3963-fabR	b4390-nadR
b0464-acrR	b1303-pspF	b1735-chbR	b2427-murR	b3021-mqsA	b3501-arsR	b3973-birA	b4393-trpR
b0487-cueR	b1323-tyrR	b1827-kdgR	b2491-hyfR	b3025-qseB	b3512-gadE	b4004-zraR	b4398-creB
b0504-allS	b1328-pgrR	b1916-sdiA	b2537-hcaR	b3060-ttdR	b3556-cspA	b4018-iclR	b4480-hdfR
b0506-allR	b1384-feaR	b1987-cbl	b2554-glrR	b3071-yqjI	b3569-xylR	b4046-zur	b4498-gatR

Boinc results

10th Sep 2015

Users	#
With credit	234
With recent credit	99
Registered in past 24 hours	0
Computers	**#**
With credit	1095
With recent credit	293
Registered in past 24 hours	1
Current GigaFLOPS	518.35

Seven days snapshot 24th Dec 2014

TN-Grid: Result summary

95988 results

Server state	# results
Inactive	0
Unsent	1080
Unknown	0
In progress	3593
Over	91315

'Over' results

Outcome	# results
---	0
Success	89536
Couldn't send	0
Computation error	1104
No reply	570
Didn't need	12
Validate error	0
Abandoned	93

'Success' results

Validate state	# results
Initial	1716
Valid	87584
Invalid	51
Workunit error - check skipped	0
Checked, but no consensus yet	10
Task was reported too late to validate	175

File Delete state	# results
Initial	1726
Ready to delete	0
Deleted	87810
Delete Error	0
Total files deleted	87810

'Client error' results

Client state	# results
Downloading	21
Processing	0
Compute error	10
Uploading	0
Done	10
Aborted by user	1063

Boinc statistics

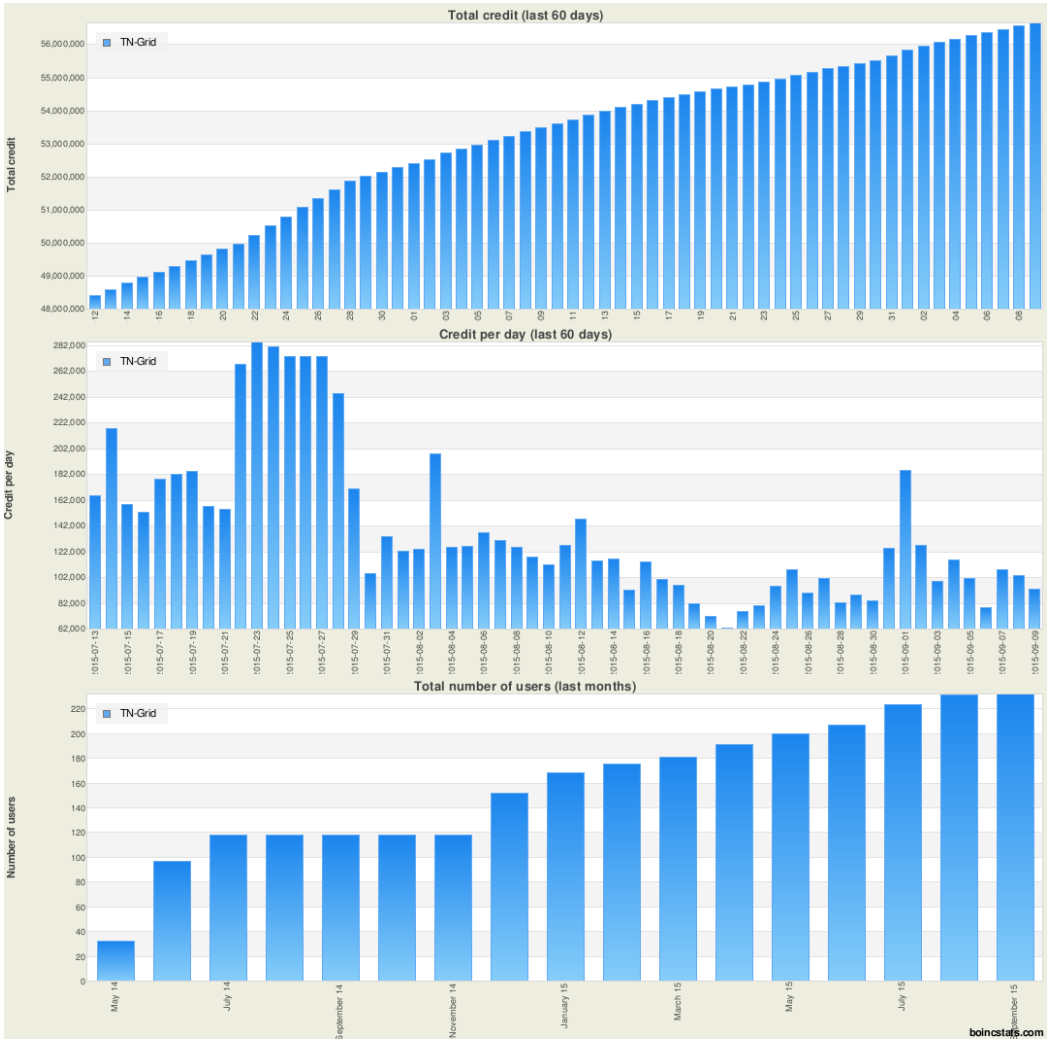
Up to 10th Sep 2015

56,693,056 cobblestones* (Boinc credits)

i5-2400 (4 cores) is capable of 3,318.2 MFlops

Equivalent to 56693056/200/3.3182/4
= 21356.86 days (~58.51 years)

*one day of work on a computer that can meet either of two benchmarks:
1,000 double-precision MFLOPS based on the Whetstone benchmark
1,000 VAX MIPS based on the Dhrystone benchmark



GENE@HOME
GEne Network Expansion

Parallel version

PC Removes graph edges as soon as detected
PC* Edge removal at the end of each (*l* level) loop
(better suited for a parallel version, easier sync)

Multithread CPU version (Intel Threading Library)
GPU version (NVidia CUDA)

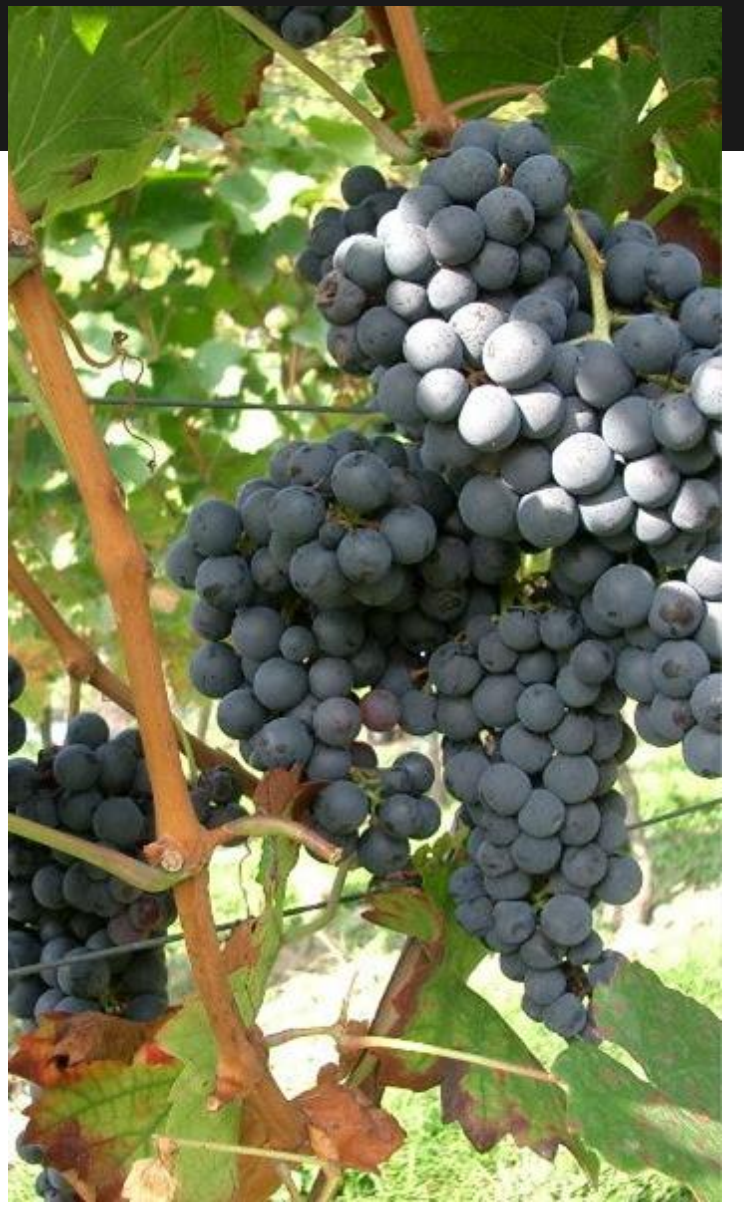
Tile size	CPU 1000	GPU 1000	CPU 2000	GPU 2000	CPU 100	GPU 100	CPU 200	GPU 200
Organism	At	At	At	At	Ec	Ec	Ec	Ec
Separation set size								
0	47	5	200	9	<1	<1	<1	<1
1	2940	600	18000	1350	<1	<1	80	3
2	1180	3100	9000	8000	90	40	890	320
3	68	100	600	220	320	100	2950	980
4	10	44	100	90	580	190	5330	1630
5	15	44	15	100	500	220	5515	2380
6	10	44	15	74	490	290	4437	3170
7			10	74	390	390	3690	4975
8			10	74	230	490	2500	6630
9					93	430	1800	8380
10							650	7020
11							230	5840
12							80	3950
13							15	2260

Future work

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

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

Regional mushrooms
(correlation / causal relationship with meteo and other data)



Questions are welcome