Time Dynamics of Protein Complexes in a Transgenic Mouse Model for Alzheimer's Disease

Fabio Cumbo

Department of Engineering
Roma Tre University

Institute for Systems Analysis and Computer Science "Antonio Ruberti"
Italian National Research Council

Master of Science Degree in Computer Engineering
2013 - 2014

Prof. Giuseppe Di Battista
Department of Engineering - Roma Tre University

Dr. Paola Bertolazzi
Dr. Giovanni Felici
Institute for Systems Analysis and Computer Science "Antonio Ruberti"
Italian National Research Council - CNR

Dr. Concettina Guerra
Georgia Tech College of Computing

Outline

Outline

jLIMITED

Goal: Design and development of an open source Java library (jLIMITED) for the analysis of "time series microarray data" to study the dynamics of "protein complexes".

Java LIbrary for MIcroarray TimE Dynamics analysis

Outline

Microarray Dataset

Microarray

A microarray is a set of DNA sequences representing the entire set of genes of an organism, arranged in a grid pattern. It's used to measure the level of expression of each gene of a biological sample.

Time series microarray dataset containing **gene expression profiles** (representing the gene activity under a particular condition) of a transgenic Alzheimer's disease affected mouse supplied by the **European Brain Research Institute** (EBRI, Rome, Italy).

class	VH											
tissues	bfb				ctx				hp			
ages	1	3	6	15	1	3	6	15	1	3	6	15

60 samples

class	AD11											
tissues	bfb				ctx				hp			
ages	1	3	6	15	1	3	6	15	1	3	6	15

60 samples

Protein Complexes Dataset

Protein complex

A group of proteins with a specific biological function.

Protein complexes were taken from the **Comprehensive Resource of Mammalian protein complexes database (CORUM)**.

Complesex dataset: **81 complexes**, of size ranging from **3 to 40**.

We traced to proteins due to a one to one correspondence between genes and proteins.

Outline

Constructing the gene co-expression matrices associated to complexes

Endoplasmic Reticulum localized multiprotein complex

	Cabp1	Dnajb11	Hsp90b1	Hspa5	P4hb	Pdia4	Ppib	Sdf2l1
Cabp1	1.000	-0.141	0.107	-0.048	-0.347	-0.134	0.012	0.294
Dnajb11		1.000	0.128	-0.453	0.216	0.751	0.527	0.390
Hsp90b1			1.000	-0.339	-0.147	0.290	0.395	0.410
Hspa5				1.000	0.041	-0.260	-0.335	-0.001
P4hb					1.000	0.340	0.018	0.241
Pdia4						1.000	0.696	0.519
Ppib							1.000	0.717
Sdf2l1								1.000

For each matrix, the average was taken over all the elements of the upper triangular matrix (excluding the diagonal).

Variation in average gene co-expression

A statistical test was applied to evaluate the significance of the change in average co-expression value within a complex. Precisely, the **paired t-student** test computes the value:

$$t = \frac{\sum_{i,j}(A_{ij} - B_{ij})}{\sqrt{\frac{N\sum_{i,j}(A_{ij}-B_{ij})^2 - (\sum_{i,j}(A_{ij}-B_{ij}))^2}{N-1}}} \tag{1}$$

where $N = \frac{n \times (n-1)}{2}$, n is the size of a complex and A_{ij} and B_{ij} are matrices values in position (i, j) related to two Pearson correlation matrices (i.e. matrix A and B respectively) in two different conditions (i.e. VH and AD11 classes or two consecutive time points for a specific class).

$$\begin{cases} \textit{if } \mid t \mid > \textit{TINV}(p) \textit{ and } (avg_A - avg_B) < 0 \rightarrow \textit{return } 1 \\ \textit{else if } \mid t \mid > \textit{TINV}(p) \textit{ and } (avg_A - avg_B) \geq 0 \rightarrow \textit{return } -1 \\ \textit{otherwise return } 0 \\ \textit{considering } p = 0.05 \ (p - \textit{value}) \end{cases} \tag{2}$$

Variation in average gene co-expression

Each element of the vector representation belongs to the set $\{-1, 0, +1\}$

Quadruplet representation of the dynamic of a complex

$$(t_1, t_3, t_6, t_{15})^{VH-AD11}_{bfb} \qquad (t_1, t_3, t_6, t_{15})^{VH-AD11}_{ctx} \qquad (t_1, t_3, t_6, t_{15})^{VH-AD11}_{hp} \tag{3}$$

Triplet representation of the dynamic of a complex

$$(t_{1-3}, t_{3-6}, t_{6-15})^{VH}_{bfb} \qquad (t_{1-3}, t_{3-6}, t_{6-15})^{VH}_{ctx} \qquad (t_{1-3}, t_{3-6}, t_{6-15})^{VH}_{hp} \tag{4}$$

$$(t_{1-3}, t_{3-6}, t_{6-15})^{AD11}_{bfb} \qquad (t_{1-3}, t_{3-6}, t_{6-15})^{AD11}_{ctx} \qquad (t_{1-3}, t_{3-6}, t_{6-15})^{AD11}_{hp} \tag{5}$$

Distance of gene co-expression matrices

Given two correlation matrices A and B corresponding to AD11 and VH samples at the same month, we adopted as an inverse measure of their similarity the **Euclidean distance** between the two matrices:

Euclidean distance

$$d = \sqrt{\frac{\sum_{i=1,j=i+1}^{N-1,N}(A_{ij} - B_{ij})^2}{N}} \tag{6}$$

where $N = \frac{n \times (n-1)}{2}$, n is the size of a complex and A_{ij} and B_{ij} are matrices values in position (i, j) related to two Pearson correlation matrices (i.e. matrix A and B respectively) in two different conditions (i.e. VH and AD11 classes or two consecutive time points for a specific class).

Distance of gene co-expression matrices

We resorted to a statistical test to provide a significance measure of d.

We determined the average distance of AD11 and VH for a collection of 1000 random generated complexes and computed the **z-score** for the distance value d of an observed complex with respect to the random complexes.

z-score

$$z = \frac{(d - \bar{x})^2}{\sigma} \tag{7}$$

where $\bar{x}$ is the average over all random samples of the distances of AD11 and VH matrices and σ denotes the standard deviation of the same distances.
z is distributed according to a normal distribution with mean 0 and variance 1.

In our test we chose a z-score of **1.9** as a threshold of significance.

Distance of gene co-expression matrices

Each element of the vector representation belongs to the set of real numbers $\mathbb{R}$

Quadruplet representation of the dynamic of a complex

$$(d_1, d_3, d_6, d_{15})^{VH-AD11}_{bfb} \quad (d_1, d_3, d_6, d_{15})^{VH-AD11}_{ctx} \quad (d_1, d_3, d_6, d_{15})^{VH-AD11}_{hp} \tag{8}$$

Triplet representation of the dynamic of a complex

$$(d_{1-3}, d_{3-6}, d_{6-15})^{VH}_{bfb} \quad (d_{1-3}, d_{3-6}, d_{6-15})^{VH}_{ctx} \quad (d_{1-3}, d_{3-6}, d_{6-15})^{VH}_{hp} \tag{9}$$

$$(d_{1-3}, d_{3-6}, d_{6-15})^{AD11}_{bfb} \quad (d_{1-3}, d_{3-6}, d_{6-15})^{AD11}_{ctx} \quad (d_{1-3}, d_{3-6}, d_{6-15})^{AD11}_{hp} \tag{10}$$

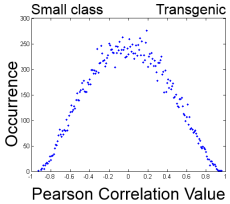
Outline

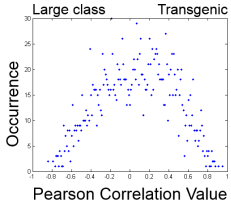
Gene correlation strength within protein complexes

We compute the Pearson correlation for each pair of genes in the same complex using the dataset of control VH mice acquired at month 1.

We divide complexes into two classes:

- **small** ($<= 10$)
- **large** (> 10)





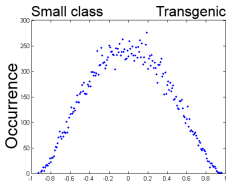
Gene correlation strength within protein complexes

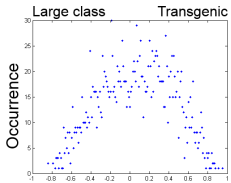
We compute the Pearson correlation for each pair of genes in the same complex using the dataset of control VH mice acquired at month 1.

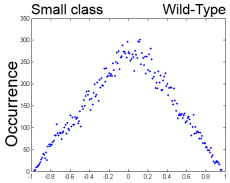
We divide complexes into two classes:

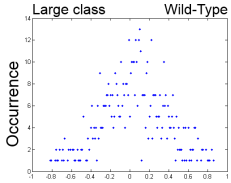
- **small** ($<= 10$)
- **large** (> 10)

We repeat this test on a public wild-type dataset.









Comparative analysis of gene co-expression values in complexes of AD11 and VH over time

We use two different measures to determine the changes of co-expression values in a complex in the two classes:

- the difference of the average co-expression between AD11 and VH using the quadruplet representation based on t-Student.
- the distance of the co-expression matrices of AD11 and VH using the quadruplet representation based on the matrix distance.

Difference of average co-expression levels in complexes

14 complexes ordered by size.

COMPLEX	Size	$t_1^{VH-AD11}$	$t_3^{VH-AD11}$	$t_6^{VH-AD11}$	$t_{15}^{VH-AD11}$
Parvulin-associated-pre-rRNP-complex	40	1	0	0	-1
20S-proteasome	14	1	0	0	0
immunoproteasome	14	1	0	0	0
B-Ksr1-MEK-MAPK-14-3-3-complex	8	0	0	-1	0
Gata1-Fog1-MeCP1-complex	8	0	1	0	0
Drosha-complex	7	1	0	0	0
BLOC-1-biogenesis-of-lysosome-related-organelles-complex	6	0	0	-1	0
Metallothionein-3-complex	6	0	0	0	-1
Brd4-Rfc-complex	5	1	0	0	0
MCM-complex	5	1	-1	0	0
Agap11-AP3-complex	4	1	0	0	0
Kif3-cadherin-catenin-complex	4	1	0	0	0
Sarcoglycan-sarcospan-syntrophin-dystrobrevin-complex	4	0	0	1	0
Wave-2-complex-Rac-activated	4	-1	0	0	0

Distance of co-expression matrices of AD11 and VH complexes

6 complexes ordered by size.

COMPLEX	size	$d_1^{VH-AD11}$	$d_3^{VH-AD11}$	$d_6^{VH-AD11}$	$d_{15}^{VH-AD11}$
immunoproteasome	14	0.26 (2.22)	0.14 (-2.45)	0.2 (-0.22)	0.16 (0.19)
Mediator-complex	7	0.29 (2.10)	0.2 (0.09)	0.16 (-1)	0.11 (1.17)
Wave-2-complex-Rac-activated	4	0.33 (1.98)	0.07 (-1.85)	0.33 (1.97)	0.36 (3.39)
p97-Ufd1-Npl4-IP3-receptor-complex	4	0.34 (2.14)	0.21 (0.31)	0.16 (-0.56)	0.2 (0.86)
Tis7-Sin3-Hdac1-Ncor1-Sap30-complex	4	0.34 (2.14)	0.2 (0.16)	0.12 (-1.16)	0.03 (-1.81)
Axin-Dvl-Gsk-Frat1-complex	4	0.28 (1.20)	0.33 (2.17)	0.13 (-1.01)	0.1 (-0.7)

Difference of average co-expression levels over time

8 complexes ordered by size resulting from the union of the previously selected complexes.

COMPLEX	Size	t^{VH}_{1-3}	t^{VH}_{3-6}	t^{VH}_{6-15}	t^{AD11}_{1-3}	t^{AD11}_{3-6}	t^{AD11}_{6-15}
Parvulin-associated-pre-rRNP-complex	40	0	0	0	-1	0	-1
20S-proteasome	14	0	0	0	0	-1	0
immunoproteasome	14	0	-1	0	-1	-1	0
Drosha-complex	7	0	1	-1	-1	0	0
BLOC-1-biogenesis-of-lysosome-related-organelles-complex-1	6	0	1	-1	0	0	0
MCM-complex	5	0	0	0	-1	0	0
Kif3-cadherin-catenin-complex	4	0	0	0	-1	0	0
Tis7-Sin3-Hdac1-Ncor1-Sap30-complex	4	0	0	0	1	0	0

Outline

Conclusions

The entire previously described method was implemented in the jLIMITED library that will be soon available on the IASI-CNR website.

All complexes obtained with our analysis are linked to the Alzheimer's disease in literature and are now the object of further investigations by the biologists of EBRI.

Part of this work is in a manuscript to be submitted for publication to an international journal.

Time Dynamics of Protein Complexes in a Transgenic Mouse Model for Alzheimer's Disease

Ivan Arisi[1], *Mara D' Onofrio*[1], *Rossella Brandi*[1], *Antonio Cattaneo*[2,4], *Paola Bertolazzi*[3], *Fabio Cumbo*[3], *Giovanni Felici*[3], *Concettina Guerra*[3,5]

1. Genomics Unit, European Brain Research Institute (EBRI), Rome, Italy
2. Neurotrophic Factors and Neurodegenerative Diseases Unit, European Brain Research Institute (EBRI), Rome, Italy
3. Institute for Systems Analysis and Computer Science "Antonio Ruberti", CNR, Rome, Italy
4. Scuola Normale Superiore, Pisa, Italy
5. College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Thanks

Thank you for your attention