Co2Vis: A Visual Analytics Tool for Mining Co-Expressed and Co-Regulated Genes Implied in HIV Infections

Amal Zine El Aabidine, Arnaud Sallaberry, Sandra Bringay, Mickaël Fabrègue, Charles-Henri Lecellier, Nhat Hai Phan, Pascal Poncelet

BioVis
2013

Co²Vis: A Visual Analytics Tool for Mining Co-Expressed and Co-Regulated Genes Implied in HIV Infections

Amal Zine El Aabidine[1,2], Arnaud Sallaberry[1,3], Sandra Bringay[1,3], Mickael Fabregue[1,4], Charles Lecellier[5], Nhat Hai Phan[1,4], Pascal Poncelet[1,2]

1 – Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier LIRMM, Montpellier, France
2 – Université Montpellier 2, Montpellier, France
3 – Université Montpellier 3, Montpellier, France
4 – IRSTEA Montpellier, UMR TETIS, Montpellier, France
5 – Institut de Génétique Moléculaire de Montpellier IGMM, Montpellier, France

Introduction

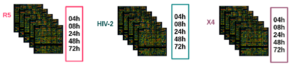
One of the key challenges in human health is the identification of **disease-causing genes** like AIDS (Acquired ImmunoDeficiency Syndrome). Numerous studies have addressed this challenge through gene expression analysis. Due to the amount of data available, processing **DNA microarrays** in a way that makes biomedical sense is still a major issue.

Statistical methods and data mining techniques play a key role in discovering previously unknown knowledge. However, applying such techniques in this context is difficult because the number of measurement points (i.e., **gene expression levels**) is much higher than the number of samples resulting in the well-known curse of dimensionality problem also called the high feature-to-sample ratio [1].

We propose a combination of **data mining** and **visual analytics** methods to identify and render **groups of genes** implied in HIV infections and sharing **common behaviors**.

Data

Temporal changes of the expression levels of about 19,000 genes on three HIV strains (HIV-2, HIV-1 : X4 and R5), were evaluated at 04h, 08h, 24h, 48h and 72h after infections.



Partially ordered patterns

This step consists in identifying **co-expressed** genes, i.e. sequential patterns of genes extracted according to their level of expression [2]:

1- Sequence processing

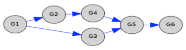
Puces	Séquences
1	<(**G2**)(**G1 G5**)(**G3**)(G4)>
2	<(**G2**)(**G1 G5**)(G4)(**G3**)>
3	<(**G2**)(G4)(**G1 G5**)(**G3**)>
4	<(G2)(G3)(G1 G5)(G4)>

2- Closed Sequential patterns

<(**G2**)(**G1 G5**)(**G3**)>
support=**3/4**

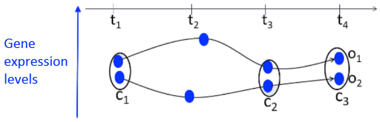
3- Partially ordered patterns

<(**G1**)(**G2**)(**G4**)(**G5**)(**G6**)>
<(**G1**)(**G3**)(**G5**)(**G6**)>



Temporal patterns

This step consists in identifying **co-regulated** genes, i.e. groups/ trajectories of genes sharing frequently a similar level of expression (incremental approach GET_MOVE [3,4]):



({o1,o2}{t1,t3,t4})

Annotation: each trajectory (e.g. the group C in the previous example) is labeled by a set of functions extracted from the gene ontology website (http://www.geneontology.org/).

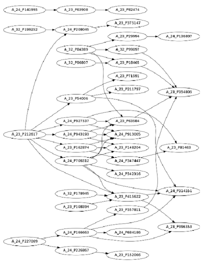
Visualization

Partially ordered patterns (on the left): DAG drawing algorithm [5]

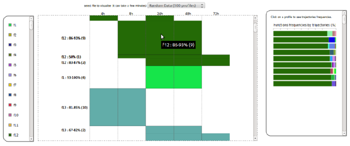
Temporal patterns (on the right):

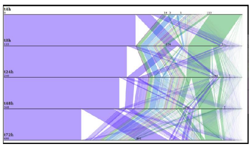
- Interactive graphic chart. The list (on the left) shows the set of functions. The diagram (on the center) shows the number of trajectories grouped by their main functions along time. The histogram (on the right) shows the distributions of the functions for each trajectory of a selected rectangle in the diagram.
- Parallel sets visualization [6] (on the bottom). Clusters of genes for each time step (computed for the temporal pattern extraction process).

Partially ordered patterns



Temporal patterns





http://www.lirmm.fr/genes_visualization/

Email: arnaud.sallaberry@lirmm.fr

[1] E.R. Dougherty. Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2:28-34, 2001.
[2] M. Fabregue, S. Bringay, P. Poncelet, M. Teisseire and B. Orsetti. Mining microarray data to predict the histological grade of a breast cancer. JBI, 44:S12-S16, 2011.
[3] H. Phan Nhat, D. Ienco, P. Poncelet and M. Teisseire. Mining Representative Movement Patterns through Compression. PAKDD, pp. 314-326, 2013.
[4] H. Phan Nhat, P. Poncelet and M. Teisseire. GET_MOVE: An Efficient and Unifying Spatio-Temporal Pattern Mining Algorithm for Moving Objects. IDA, pp. 276-288, 2012.
[5] E. R. Gansner, E. Koutsofios, S. C. North, and K.-P. Vo. A Technique for Drawing Directed Graphs. IEEE TSE, 19(3):214–230, 1993.
[6] F. Bendix and R. Kosara and H. Hauser. Parallel sets: A visual analysis of categorical data. InfoVis, pp 133-140, 2005.