Independent component analysis to investigate microRNA role in Ewing sarcoma

Martignetti L[(1)], Zinovyev A[(1)], Stoll G[(1)], Tirode F[(2)], Laud-Duval K[(2)], Delattre O[(2)] and Barillot E[(1)]

[(1)] Bioinformatics, biostatistics, epidemiology and computational systems biology of cancer - Inserm Unit 900 / Ecoles des Mines / Institut Curie;
[(2)] Genetics and Biology of Cancers - Inserm Unit 830 / Institut Curie

Motivation

The Ewing tumor is characterized by a malignant genomic translocation and appearance of a chimeric gene EWS/FLI-1 with transcription factor activity leading to uncontrolled cell growth [1]. Since the expression of the EWS-FLI1 gene alone can change cell phenotype from normal to tumorigenic in fully reversible and controlled manner, this is an excellent system for understanding the complex picture of deregulations happening in cancer cells. The availability of high-throughput data for Ewing patients, including transcriptome and miRNAs expression data, allows us to infer new regulatory interactions involved in the EWS/FLI-1 network and to elucidate the impact of the miRNAs activity. However, exploitation of large amount of data generated by high-throughput miRNAs expression profiles is challenging. MiRNA expression is controlled by many cellular variables and the global picture of their expression profiles is complex and noisy. We expect each miRNA to be influenced by several factors, like different transcription regulators, cellular processes and biological responses. Independent component analysis (ICA) is a statistical and computational technique for revealing hidden factors that underlie sets of measurements or signals. This approach has been already successfully applied to analyse gene expression data [2]. Thus we extend its application to miRNA expression data. It relies on the idea of a combinatorial control, describing the expression level of miRNAs as linear functions of common hidden variables. According to the ICA model, the hidden variables exert linear influences on the miRNAs with no statistical dependences between them. Each independent component defines groups of commonly influenced miRNAs and we expect to find among the independent components some which are of biological significance. Differently from a standard clustering approach, assigning each miRNA to one cluster with correlated expression pattern, we expect each miRNA can participate, to varying degrees, in many independent pattern of covariation. Once components derived by ICA have been identified, this analysis also reveals those individual miRNAs that have the strongest contribution to each component.

Methods

We apply ICA to a miRNA dataset containing the expression profile of 740 miRNAs for 20 Ewing sarcoma patients. We use the FastICA algorithm [3] and we extimate 10 robust and reliable independent components by running it 50 times and combining information from several runs by the Icasso tool [4]. To establish a possible biological relevance of miRNA groups identified by ICA, we investigate different hypothesis. Firstly, deregulated miRNAs can be located in genetically altered regions associated with cancers. We check for chromosome distribution of miRNA belonging to each component and possible highly recurrent genomic loci. Following a second hypotesis, we investigate whether miRNA groups are enriched for members of the same miRNA family that have highly similar binding sites and may coordinately regulate common target genes. Furthermore, we test if ICA is able to identify novel patterns of coregulated miRNAs significantly associated with clinical outcome (metastatic versus non metastatic). The defined miRNA components have been analyzed by using univariate t test, searching for a miRNA signature that characterizes metastatic and non-metastatic subgroups.

Results

The application of ICA allows us to reduce the complexity of miRNA expression profiles to a small number of regulatory variables and to group miRNAs that are strongly related to the same component. Comparing the 10 identified miRNA groups with miRNA families annotated in mirBase 9.2, any component shows significant enrichment for members of the same family, according to the hypergeometric test. On the contrary, the analysis of chromosomal distribution clearly associates the most robust component to the miRNA cluster in the human Dlk1-Gtl2 domain at 14q32. This domain is known to be expressed in a large non-coding transcriptional unit which is altered in human ovarian cancer [5]. Moreover, when trying to associate our miRNA

groups to the clinical outcome of Ewing disease, miRNAs of Dlk1-Gtl2 domain associated to the most robust component, taken toghether, create a significant expression signature able to distinguish malignant from non-malignant Ewing patients. These findings confirm the potential of ICA to extract information from large and complex miRNA expression data. The analysis supports promising indications that miRNAs at Dlk1-Gtl2 domain can be involved in Ewing sarcoma and in the malignancy of the disease.

References

[1] Delattre O et al., N Engl J Med, 1994
[2] Martoglio AM et al, Bioinformatics, 2002
[3] Hyvarinen A, IEEE Trans Neural Netw., 1999
[4] Himberg J, IEEE Neural Netw Sign Processing, 2003
[5] Zhang et al, PNAS, 2008

Contact : loredana.martignetti@curie.fr