DISCOCLINI: a system for Biomarkers Discovery in Medical Functional Genomics data

Arriel Benis, Mélanie Courtine, Alain Venot

Laboratoire d'Informatique Médicale et Bioinformatique (LIM&BIO EA 3969), UFR SMBH, Université Paris 13, Bobigny, France

Abstract and Objective

One of the difficult problems of data interpretation in the analysis of cDNA chips is that the number of expression levels per chip is very high compared to the number of chips. Our method, DISCOCLINI, consists in calculating all correlations in data and reformulating them to easily visualize variation patterns.

Keywords:

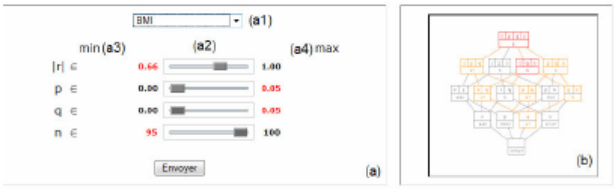
Gene expression, Biological and clinical data, Data mining, Knowledge discovery, Correlation, Visualization.

Introduction

Data Mining is an emerging area in Medical Informatics research field. It consists in extracting knowledge from large databases in order to assist physicians. Nowadays, clinical research protocols are no longer limited to collect only medical data, but they are also regarding to other kinds of data such as genomic data from cDNA microarrays. The data analysis has to take into account all these data in despite of their different nature and their relative quality. Currently, the "classical" approaches commonly used by biologists in this context simply explore a tiny part of the data mainly using a priori selections.

Method

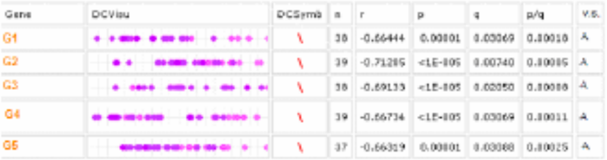
Our system, DiscoClini, is based on a data workflow adapted to data that we want to deal with (bioclinical data versus cDNA microarrays data). The analysis method consists mainly in calculating the Spearman's rank correlation coefficient, for all the bioclinical vs. cDNA chips datasets. Then, all these results (around 400 000) need to be presented to biologist experts in an easy way for interpretation. This is done in two steps. Firstly (Figure 1(a)) a Visual Data Mining interface allows the user to choose which bioclinical parameter he wants to study and he dynamically defines the thresholds of the values of the correlation coefficient, the significance, the fault discovery rate and the number of individuals involved in the relations. Secondly (Figure 1(b)), a Hasse diagram of all the relationships calculated and corresponding to the chosen threshold is built. When we select one of its nodes, we obtain the description of all selected relationships (Figure 2), the most relevant one according to the expert.



BMI (a1)

	min (a3)	(a2)	(a4) max
|r| ∈	0.66		1.00
p ∈	0.00		0.05
q ∈	0.00		0.05
n ∈	95		100

Figure 1-Screenshot of the Visual Data Mining interface.



Gene	DCVisu	DCSymb	n	r	p	q	p/q	V.S.
G1		\	38	-0.66444	0.00001	0.03069	0.00018	A
G2		\	39	-0.71285	<1E-005	0.00740	0.00085	A
G3		\	38	-0.69133	<1E-005	0.02050	0.00088	A
G4		\	39	-0.66734	<1E-005	0.03069	0.00011	A
G5		\	37	-0.66319	0.00001	0.03088	0.00025	A

Figure 2-Example of visualization of relationships using DcVisu obtained after selection of one of the nodes of the Hasse diagram (Figure 1(b)).

Results and Conclusion

Experiments related to researches in obesity medicine have been done [1,2,3]. They allowed us to validate our Data Mining process, step by step, and to discover "potential" biomarkers. Evaluation of use and usability has shown the benefits of the system as a whole.

References

[1] Viguerie N, Clément K, Barbe P, Courtine M, Benis A, Larrouy D, Hanczar B, Pelloux V, Poitou C, Khalfallah Y, Barsh GS, Thalamas C, Zucker JD, Langin D. In vivo epinephrine-mediated regulation of gene expression in human skeletal muscle. J Clin Endocrinol Metab, 2004: 89(5), 2000–14. 0021.

[2] Clément K., Viguerie N, Poitou C, Carette C, Pelloux V, Curat C, Sicard A, Rome S, Benis A, Zucker J, Vidal H, Laville M, Barsh G, Basdevant A, Stich V, Cancello R, Langin D. Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. FASEB J, 2004: 18(14), 1657–1669.

[3] Taleb S., Lacasa D, Bastard J, Poitou C, Cancello R, Pelloux V, Viguerie N, Benis A, Zucker J, Bouillot J, Coussieu, C Basdevant A, Langin D, Clément K. Cathepsin s,

a novel biomarker of adiposity: relevance to atherogenesis. FASEB J, 2005: 19(11), 1540–2.