Identification of dilated cardiomyopathy signature genes through gene expression and network data integration

A Camargo, F Azuaje*

University of Ulster at Jordanstown, School of Computing and Mathematics, Shore Road, Newtownabbey, Co. Antrim, BT37 0QB, Northern Ireland, UK.

av.camargo@ulster.ac.uk
* Corresponding author: fj.azuaje@ulster.ac.uk

Background

Dilated cardiomyopathy (DCM) is a leading cause of heart failure (HF) and cardiac transplantations in Western countries (Barth et al., 2006, AHA, 2007). Gene expression studies have applied different methodologies to gain insights into the aetiology of this disease (King et al., 2005; Barth et al., 2006; Wittchen et al., 2007). However, it remains uncertain whether the integration of those independent data sets may improve systems-level knowledge and support potential clinical applications. Moreover, there are concerns in connection to the reproducibility of prediction results. The main hypothesis of this study is that the integration of publicly available data and information sources (Bijnens et al., 2006), may support the identification of potentially relevant, reproducible disease biomarkers. This study identified disease signature genes by analysing the expression profile of human DCM and non-DCM samples derived from different studies. First, two heterogeneous, lab-independent data sets with DCM and non-DCM samples were integrated. Second, gene expression analyses identified significantly differentiated genes among the two classes. Third, the Nearest Shrunken Centroid algorithm identified class-distinguishing genes, whose expression profiles can be used to identify DCM samples automatically. Results regarding the class-distinguishing genes and the differentially expressed genes' expression profiles were analysed in a third, independent data set. To analyse significant biological responses and relationships, class- distinguishing genes and differentially expressed genes, as well as their corresponding associated KEGG and Reactome pathways were mapped onto a Protein-Protein Interaction (PPI) network. This network, which is one of the by-products of this research, represents a global human heart failure survey of PPIs obtained from the Human Protein Reference Database (HPRD) (Peri et al., 2003). This sequence of analyses identified potentially novel DCM signature genes, whose class prediction results based on expression patterns could be replicated using an independent testing dataset. The integration of expression profiles and PPI network analysis identified genes that are involved in several biological pathways, which may be relevant to the disease evolution. This study demonstrates that the integration of microarray data and other sources of functional information may support the generation of testable hypotheses about potentially novel disease bio-markers and drug targets for DCM in humans.

Results

Gene expression analysis

Three microarray data sets generated by independent studies in human heart failure were obtained from the GEO (Gene expression Omnibus), accession numbers GDS2205, GDS2206 and GSE2143. The latter (testing dataset) was used to reproduce prediction results obtained from the integration of the first two data sets. In GDS2206 and GSE2143 there were 40 samples in total: 20 corresponding to Non-DCM and 20 to DCM samples. After data transformation and normalisation, a total of 2482 probe sets were selected for further analyses. A first analysis phase applied the significance analysis of microarray method (SAM) (Tusher et al., 2001) to identify significantly differentiated genes between Non-DCM and DCM samples. This test identified 97 significantly differentiated genes: 6 down-regulated and 91 up-regulated in DCM. Of them, genes AP3M2 (adaptor-related protein complex 3), EEF1A1 (eukaryotic translation elongation factor 1 alpha 1), LAMB1 (laminin, beta 1) and ODC1 (ornithine decarboxylase 1) were also significantly differentiated when expression analyses were independently carried out on the data sets by the authors of the original studies (Barth et al., 2006). A second analysis phase applied Nearest Shrunken Centroid (Tibshirani et al., 2002) to identify class-distinguishing genes, whose expression patterns may be used to differentiate Non-DCM from DCM samples. This analysis identified 15 class-distinguishing genes, which provided the basis for overall classification accuracy equal to 90.0% (based on 10-fold cross-validation). The set of optimal inputs to this classifier comprised the following genes: HTRA1, LAPTM4B, ANAPC13, DLD, CSDE1, CAT, NPPA, SNX3, UBB, HSP90AB1, TAF7, CAP2, TMEM66, SKP1A and ODC1. These genes also ranked the highest in the SAM (FDR <0.001).

In order to support the assessment of the potential biological relevance of these results, classification analyses were replicated, using the class-distinguishing genes previously identified, on an independent, testing data set. This analysis showed that class-distinguishing genes were also down-regulated in DCM. Genes LAPTM4B and NPPA were the highest ranked by SAM (FDR <0.001, folding change > 2). Figure 1 depicts the class-distinguishing genes' expression profiles observed in the testing data set. Expression profile corresponding to LAPTM4B is illustrated at the bottom of the graph.

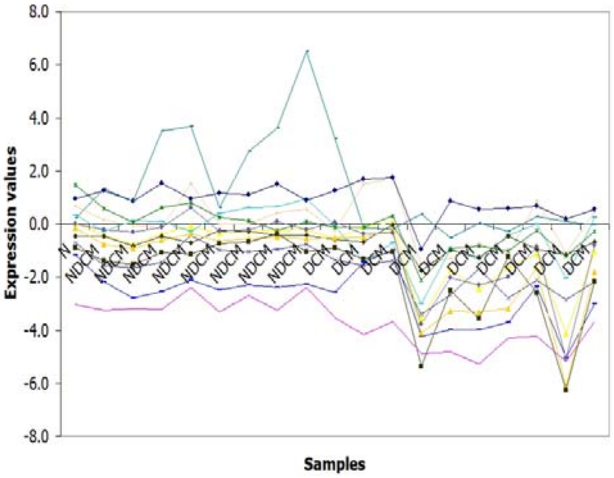
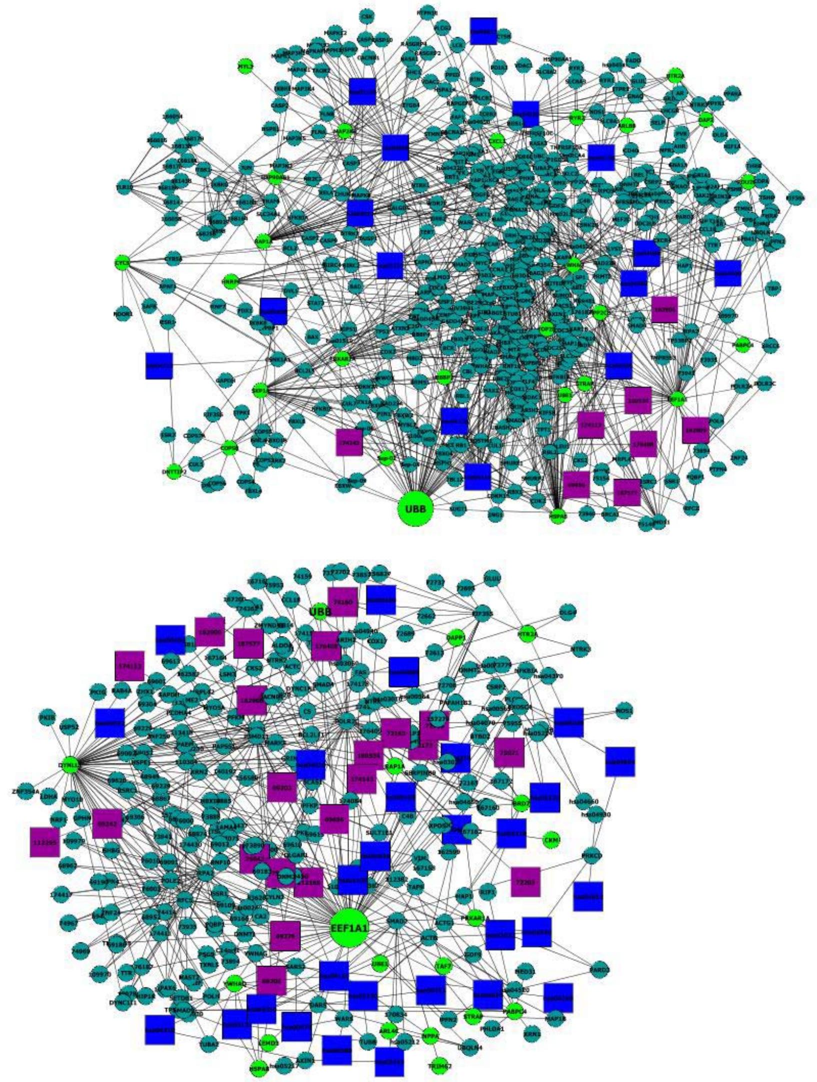


Figure 1 Expression values (Log2) of class-distinguishing genes as reproduced in testing data set. Graph depicts expression distribution of class- distinguishing genes across N-DCM and DCM samples.



(a)

(b)

Figure 2 PPI network composed of proteins encoded by significantly differentiated or class-distinguishing genes (green nodes), Protein interaction partners (teal nodes), KEGG pathways (dark blue nodes) and Reactome (purple nodes) pathways. Circles represent single proteins. Boxes represent functional pathways. (a) UBB PPI network (b) EEF1A1 PPI network.

PPI network analysis

The second part of the study analysed gene expression profiles in the context of a human HF-related PPI network. We applied a slightly similar approach as Lu et al. (2007), who analysed allergic response in experimental asthma by linking gene expression data versus the topology of a PPI network. The graphical representation of our DCM network (Fig. 2) depicts nodes (proteins) colour-coded as follows. Nodes in green encode significantly differentiated, which included class-distinguishing genes. Their interacting partner proteins were represented by nodes in teal. KEGG functional pathways were represented by boxes in dark blue. Reactome functional pathways were represented by boxes in purple. The PPI network analysis showed that nodes representing proteins encoded by class-distinguishing genes (e.g. DLD, CAT, UBB, HSP90AB1, TF7, SKP1A or NPPA) interact, directly or indirectly, with a number

KEGG and Reactome pathways, which have been previously suggested as relevant processes in the development of human HF, e.g. MAPK signalling, apoptosis, p53-Independent DNA Damage Response, p53-Independent G1/S DNA damage checkpoint pathways. To facilitate the interpretation of the map, nodes representing the proteins encoded by the class-distinguishing genes are depicted with a bigger size. A closer view of the region occupied by UBB (ubiquitin B) in this network is shown in Figure 2 (a). The topological analysis of the network identified highly connected (> 30) nodes, some of which represented either significantly differentiated or class-distinguishing genes. For example, genes such as EEF1A1 (65 connections), UBB (53 connections) or SKP1A (32 connections) encode proteins that are strongly connected in the network. What is more, proteins encoded by these genes interacted with known HF-associated genes, such as TP53, AKT1 (this one is involved in more that 20 functional pathways) and STAT3. Close-up views of the network region involving EEF1A1 is shown in Figure 2 (b). Regarding the gene LAPTM4B, neither interacting partner proteins nor functional pathways were found to be associated with the protein encoded by this gene. However, according to Entrez, LAPTM4B has an active role in disease progression of malignant cells and is involved in cell proliferation and multi-drug resistance. Thus, the functional roles of this gene and its association with the development of HF motivate further experimental analyses.

Conclusions

The main hypothesis of this study was that the integration of independent microarray data sets could lead to the identification of signatures genes in human DCM. This required rigorous data pre-processing procedures to assure the integration of data sets obtained from different studies on DCM. This was followed by data analyses to select class-distinguishing genes. These genes show expression profiles that may be used to aid in the identification of DCM samples. In order to support the assessment of the potential biological relevance of these results, analyses were replicated, using the class-distinguishing genes previously identified, on an independent data set. In addition, we built a PPI interaction network to analyse biological responses and relationships of these DCM predictor genes. This network-based analysis suggested that proteins encoded by these genes are involved in several, specific biological pathways such as apoptosis, DNA repair and checkpoint, focal adhesion or cytokine-cytokine receptor interaction, which are known to be involved in the development of HF. Moreover, new potentially relevant pathways have been identified. The HF PPI assembled here represents by itself a useful compendium of the current status of human HF-relevant interactions. The network also discloses curated biological interactions, which may give additional insights into the relationship gene expression and functional responses through PPI.

This investigation suggests that the significant genes identified may be reliable, reproducible DCM predictors. This investigation also showed how the large-scale, computational integration of independent microarray data sets, functional networks, functional annotation databases and published literature may support the identification and assessment of potential therapeutic targets. In this particular case, a set of representative disease-related genes were detected, which are suggested as testable hypotheses in relation to their roles in DCM progression. Finally, given the results obtained through the PPI network assessment, we suggest that the exploratory integrative analysis of differential gene expression and PPI network analysis may facilitate a better understanding of functional roles and identification of potential therapeutic targets in human HF.

Acknowledgement

This work was supported in part by a grant from EU-FP6, CARDIOWORKBENCH project, to FA.

References

American Heart Association (AHA). 2007. Heart Disease and Stroke Statistics — 2007 Update. American Heart Association.

Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, Kääb S, Kreuzer E, Steinbeck G, Mansmann U, Poustka A, Nabauer M, Sültmann H. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. J Am Coll Cardiol. 2006, 48(8):1610-7.

Bijnens APJJ, Lutgens E, Ayoubi T, Kuiper J, Horrevoets AJ, Daemen MJAP. Genome-Wide Expression Studies of Atherosclerosis: Critical Issues in Methodology, Analyses, Interpretation of Transcriptomics Data. Arterioscler Thromb Vasc Biol 2006, 26: 1226-1235.

King JY, Ferrara R, Tabibiazar R, Spin JM, Chen MM, Kuchinsky A, Vailaya A, Kincaid R, Tsalenko A, Deng DX, Connolly A, Zhang P, Yang E, Watt C, Yakhini Z, Ben-Dor A, Adler A, Bruhn L, Tsao P, Quertermous T, Ashley EA. Pathway analysis of coronary atherosclerosis. Physiol Genomics. 2005, 23(1):103-108.

Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, Conte JV, Tomaselli G, Garcia JG, Hare JM. Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. Physiol Genomics. 2005, 3:99-307.

Lu X, Jain VV, Finn PW, Perkins DL. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Mol Syst Biol. 2007, 3:98.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003, 13(10):2363-71.

Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 2002. 99, 6567-6572.

Tusher VG, Tibshirani R, Chu G. Significance analyses of microarrays applied to the ionizing radiation response. PNAS 2001, 98(9): 5116-5121.

Wittchen F, Suckau L, Witt H, Skurk C, Lassner D, Fechner H, Sipo I, Ungethüm U, Ruiz P, Pauschinger M, Tschope C, Rauch U, Kühl U, Schultheiss HP, Poller W. Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets. J Mol Med. 2007, 85(3):253-67.