Network-based analysis of stem cells differentiation

Francesca Mulas[1]✉, Lan Zagar[2], Blaz Zupan[1], Riccardo Bellazzi[1,3]

[1]Centre for Tissue Engineering, University of Pavia, Pavia, Italy
[2]Faculty of Computer Science, University of Ljubljana, Ljubljana, Slovenia
[3]Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Pavia, Italy

Motivation and Objectives

Understanding the real developmental stage of reprogrammed stem cells is still a demanding task for researchers in regenerative medicine. In fact, developmental biology is in need of methods that would accurately predict the developmental stage reached by cells cultured in non standard conditions, such as the induced Pluripotent Stem Cells (iPSCs). Bioinformatics approaches would be extremely useful for assessing the pluripotency status of the cells, and thus, their potential use in the clinics for repairing malfunctioning tissues and organs.

To this aim, several works have recently demonstrated the utility of applying dimensionality reduction techniques to genome wide expression data (Aiba et al, 2009). As we have shown (Zagar et al, 2011), these methods can be successfully used to predict the developmental stage of cells by mapping their transcriptional profile to a one-dimensional ruler, that we named differentiation scale. The proposed approach was also useful for identifying reduced subsets of genes that drive each developmental stage (Mulas et al, 2012).

A crucial issue in this field is the integration of the findings extracted from the data with the available knowledge coming from biological databases. For instance, genes that are surrounded by a high number of selected genes in the protein-protein interaction networks should be included in the analysis (Nitsch et al, 2010). In this work, we developed a network-based pipeline for analyzing temporal gene expression data coming from embryonic stem cells differentiation. The results highlighted the transcriptional changes occurring during development and allowed identifying known markers as well as novel gene candidates potentially involved in the regulation of stem cell differentiation.

Methods

Stem cell differentiation is characterized by an intense transcription activity where a number of transcription factors regulates the gene expression of specific targets (Zagar et al., 2011). These transcriptional changes can be observed by analyzing the genome-wide expression profiles of m genes at n different samples along differentiation. Principal Component Analysis (PCA) may be used to assign a real number p(s) to a sample s based on its expression profile. The result of this inference is a set of real numbers that can be placed in a 1D ruler, the differentiation scale. To construct a more robust predictive model, we combined the expression values from six data sets on embryonic stem cells differentiation provided by Gene Expression Omnibus with a meta-analysis method named Merging. Thanks to this approach, we obtained an integrated scale where a new uncharacterized sample can be projected to uncover its real stage of development with respect to the normal dynamics.
Reduced subsets of genes specifically activated in different stages of differentiation were identified by means of a novel gene selection procedure that assigns to each gene a score proportional to its PCA-inferred weight in the stage s and its expression value.

In order to obtain a complete picture of the gene transcription in each stage, we exploited the biological knowledge available through the STRING database (Szklarczyk et al., 2011). STRING imports and combines data gathered from heterogeneous sources to provide information about known and predicted protein-protein associations. A confidence score is also assigned to each predicted association.

In this work, we developed and analyzed a set of STRING-based networks, one for each stage of development, by applying the following procedure:

1. Network building. For each stage-specific list of genes, we mapped the gene symbols to their corresponding protein ids provided by the UNIPROT database. This step allowed retrieving the protein associations predicted by STRING, that we used to build a set of gene networks. A pair of genes in each network was connected if the confidence score of their proteins association in STRING exceeded a selected global threshold.

2. Topological analysis. The developmental stages measured in the considered experiments were grouped into three phases according to the clusters of projections that we observed in the differentiation scale. We analyzed the similarity of the networks obtained for each phase in terms of their topological properties. First, in order to take into account the number of common gene links in the networks, we computed the median jaccard index between networks of different phases. Moreover, to quantitatively characterize the importance of the nodes in our networks, we considered a topological index known as betweenness centrality. This measure is defined as the number of shortest paths that go through a considered node, and represents the influence of that node in the flow of information within the network. Nodes with a high betweenness typically make possible the communications in the network among clusters of nodes characterized by high internal connectivity. The betweenness values of the genes present in each phase were used to compare the different developmental phases and to identify the most relevant genes in each network.

3. Biological analysis. The expanded lists of genes obtained with STRING were further analyzed in light of the knowledge on developmental processes reported in the literature. Different works have recently pointed out the existence of a set of specific genes that are responsible for a particular pluripotency status of the cell (Zuccotti et al., 2011). First, in order to evaluate the effect of the network-based procedure on the gene selection, we compared the list of genes retrieved with our approach with the known markers.

The analysis then focused on identifying the most biologically significant genes for each differentiation phase. The importance of a gene in a developmental stage is represented by its role in the transcriptional regulatory circuitry of the process and is best quantified by its betweenness centrality value. We therefore looked for sets of significant genes among the bottlenecks of the developed networks. Phase-specific important genes were identified applying a selection procedure based on a global threshold value. For each phase, we determined a list of characterizing genes by extracting those that were contained in more than the 60% of the total number of networks present in the phase. All genes included in at least one phase-specific list were assigned the median of all their betweenness centrality values. We then considered the distribution of such median values and computed the 95th percentile, which was assumed as the threshold. Finally, from each phase-specific gene list all genes whose betweenness centrality value exceeded the threshold were extracted.

Results and Discussion

TThe results of the presented procedure confirmed that a transcriptional wave is active during differentiation and influences the topological properties of the networks. While the analysis of the Jaccard index showed no significant phase characterization in terms of common edges, a phase comparison based on betweenness similarity highlighted instead a distance between the first phase and the last stages (Table1).

The biological analysis, i.e. the study of the expanded lists of genes in light of knowledge on developmental processes reported in the literature, identified 53 known markers included in at least one network. In particular, a number of known key genes retrieved with this method, such as Pou5f1, Nanog, Klf4, Sox2, were not previously selected by the data-driven procedure based on their expression profiles. This result confirmed how network-based approaches can integrate experimental findings contributing to the identification of significant genes, whose importance is due to the crucial role they play in the regulation of the global network.

Gene characterization based on betweenness centrality selected a higher number of genes (24) in the first phase if compared to the others (11 and 9 genes for the second and the last phase, respectively), confirming a distinction of the early stages, where the majority of transcriptional changes are known to occur. Known markers as well as novel yet uncharacterized genes were identified. Starting from these results, future experiments will be focused on the application of network-based prioritization procedures, that would help to automatically retrieve the most significant genes in the networks.

Table1: Topological similarity of the networks for the three phases of differentiation.

Phases	1:2	1:3	2:3
Jaccard	0,78	0,8	0,75
Betweenness	0,58	0,58	0,91

Acknowledgements

This work was supported by the Fondazione Cariplo grant (2008–2006) "Bioinformatics for Tissue Engineering: Creation of an International Research Group" and by EU FP7 project "CARE-MI" and grants from Slovenian Research Agency (P2-0209, J2-9699, L2-1112). Camilla Colombo is gratefully acknowledged for her help in software development.

References

1. Aiba K et al (2009) Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells, DNA Res 16:73-80. doi:10.1093/dnares/dsn035
2. Mulas F et al (2012) Supporting Regenerative Medicine by Integrative Dimensionality Reduction, Methods Inf Med. 51(4). doi: 10.3414/ME11-02-0045
3. Nitsch D et al (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics 11:460. doi: 10.1186/1471-2105-11-460
4. Szklarczyk D et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561-8. doi: 10.1093/nar/gkq973
5. Zagar L et al (2011) Stage prediction of embryonic stem cell differentiation from genome-wide expression data. Bioinformatics 27(18):2546-53. doi: 10.1093/bioinformatics/btr422
6. Zuccotti M et al (2011) Gatekeeper of pluripotency: a common Oct4 transcriptional network operates in mouse eggs and embryonic stem cells. BMC Genomics 12:1-13. doi: 10.1186/1471-2164-12-345