Prediction of cellular developmental stage from whole-transcriptome expression profiles

Mulas F[1], Zagar L[2], Sacchi L[3], Garagna S[4], Shaulsky G[5], Zuccotti M[6], Zupan B[2,5,1], Bellazzi R[3]

Motivation

Recent studies on model organisms have shown that a variety of developmental processes, including meiosis or stem cells differentiation, are governed by successive waves of gene transcriptions. In these studies, RNA samples collected at different stages of cell development are often used to characterize genes and determine their role in development. An alternative use of high-throughput experiments is to focus on individual samples. The transcription profile of a biological sample can encode the state of the organism or cell culture and provide means for assessing the developmental stage of a cell, that is usually determined with morphological or other observable markers. We propose to use the information encoded in the whole-transcriptome profiles to construct stage prediction models, called development nomograms, that can reliably predict a cell's developmental stage. A development nomogram is essentially a graphical device – a rule – whose scale reflect developmental stages and expose the dynamics of the observed process. This rule-based visualization is associated with a model that takes the transcriptional phenotype of a sample, projects it on the rule, and predicts the developmental stage of the sample. In stem cell differentiation experiments the positions of samples along the nomogram can be used as a scale for revealing the developmental potency of cells.

Methods

We have constructed development nomograms from whole-genome transcription profiles of cells observed at various stages of development, corresponding to different time points. The approach employed a combination of gene subset selection techniques and dimension reduction methods. To reduce the computational cost of the analysis and to focus our inference on the genes that best characterize the various stages of cellular development, we experimented two different methods proposed in the literature for gene subset selection in time-

[1] Centre for Tissue Engineering, University of Pavia, Pavia, Italy [2] Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia [3] Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia, Italy [4] Dipartimento di Biologia Animale, Laboratorio di Biologia dello Sviluppo, University of Pavia, Pavia, Italy [5] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA [6] Sezione di Istologia ed Embriologia, Dipartimento di Medicina Sperimentale, University of Parma, Parma, Italy

sequence data. Samples represented with expression values of the selected genes were then projected to a 1-dimensional space, either via unsupervised methods or by additionally using the stage information. In particular, we used principal component analysis (PCA) and partial least square regression (PLS) to project each time point to the nomogram. PCA has the peculiarity that can infer both the scale and the projection mechanism from the data; on the other hand, predictions of PLS may be more accurate due to supervision. We then tested the predictive power of the development nomograms either within the same data set or on different data sets and different cell lines. To evaluate the predictions when the samples are projected within the same data set, we used a leave-pair-out evaluation scheme that removes two developmental stages, infers the nomogram from the remaining training set, and then tests the prediction on the samples from the two stages that were left out (A and B in the figure). The score that we used to evaluate the accuracy of a nomogram is the AUC, represented by the proportion of sample pairs for which the inferred order of the stages corresponds to the original order of samples.

Results

In a series of experiments comprising of 14 data sets from the Gene Expression Omnibus repository, we demonstrated that the proposed approach is robust and has excellent prediction ability. The differences between projection methods were not statistically significant and the majority of the AUC scores for any of the inference methods we have implemented are close to 0.9, a score indicating a very high quality of predictions. With such scores, we can conclude that development nomograms can accurately predict developmental stages within a chosen cell line and, more surprisingly, also across different cell lines. Moreover, the staging is easy to interpret by biologists, and the visualization underlines the dynamics of the changes such that stages that are phenotypically different appear farther apart on the nomogram.

Contact e-mail

francesca.mulas@unipv.it

Image

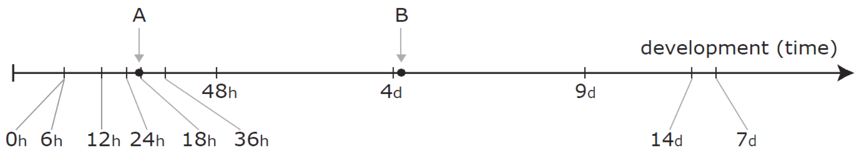

Figure 1 - Development nomogram and stage prediction for mouse embryonic stem cell differentiation