BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–8

# Stage prediction of embryonic stem cell differentiation from genome-wide expression data

Lan Zagar [1,†], Francesca Mulas [2,†], Silvia Garagna [3], Maurizio Zuccotti [4], Riccardo Bellazzi [5,2] and Blaz Zupan [1,6,2*]

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

Received on March 1, 2011; revised on N/A; accepted on N/A

Associate Editor: N/A

**ABSTRACT**

**Motivation:** The developmental stage of a cell can be determined by cellular morphology or various other observable indicators. Such classical markers could be complemented with modern surrogates, like whole-genome transcription profiles that can encode the state of the entire organism and provide increased quantitative resolution. Recent findings suggest that such profiles provide sufficient information to reliably predict the cells developmental stage.

**Results:** We use whole-genome transcription data and several data projection methods to infer differentiation stage prediction models for embryonic cells. Given a transcription profile of an uncharacterized cell, these models can then predict its developmental stage. In a series of experiments comprising 14 data sets from the Gene Expression Omnibus we demonstrate that the approach is robust and has excellent prediction ability both within a specific cell line and across different cell lines.

**Availability:** Model inference and computational evaluation procedures in the form of Python scripts as well as the used data sets are available at `http://www.biolab.si/supp/stagerank`.

**Contact:** blaz.zupan@fri.uni-lj.si

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Embryonic stem cells (ESCs) and other pluripotent cell types are increasingly being studied for their potential therapeutic use in regenerative medicine (Bhattacharya *et al.*, 2009). ESCs are isolated from the inner cell mass of the blastocyst, they replicate indefinitely, maintaining pluripotent characteristics and may differentiate in vitro to most of the somatic cell types present in the adult. While the stages of ESC differentiation into a specific cell type have been broadly identified, numerous aspects of this process remain unknown or difficult to interpret. Differentiation is a complex, multiple-steps process that presents a non linear progression within a cell population. The stem status is not lost immediately, but it gradually decreases. This is true particularly at the very beginning when differentiation is induced (*e.g.*, either by an internal or external signal) and cells own a heterogeneous status of differentiation being a mixture of diverse developmental stages.

ESCs as well as embryonic carcinoma and induced pluripotent stem cells (Müller *et al.*, 2008) own common and specific molecular signatures that define their pluripotent status. When differentiation is induced, this molecular signature is gradually lost in favour of one that defines a more differentiated type of cellular identity. Novershtern *et al.* (2011) demostrated that this cellular transition is due to a large number of transcription factors whose expression changes across different hematopoietic states. Other recent studies of various developmental processes have shown that they are governed by transcriptional programs in which genes are regulated in successive waves of transcriptions that mark the stages of differentiation (Mata *et al.*, 2002; Wagner *et al.*, 2005; Bhattacharya *et al.*, 2009; Van Driessche *et al.*, 2002; Ravasi *et al.*, 2010; Cannistraci *et al.*, 2010; Neri *et al.*, 2011). Thus, cell's transcriptional profiles could be used as whole-genome markers of differentiation.

In this work we present models that, given the transcription profile of a cell, predict its differentiation stage. Differentiation is a continuous process, and for interpretation it could be convenient if the model would map whole-genome transcription profiles to a one-dimensional projection. In this paper, we refer to this projection as a *differentiation scale*, and evaluate it on the basis of preservation of the order of data points with respect to the staging of differentiation. Projection of differentiation landmarks on this scale may further expose the dynamics of the observed process. To this end, we investigate the utility of various state-of-the-art data transformation approaches and their predictive accuracy in a systematic evaluation on 14 publicly available cell differentiation data sets from mouse, rat and human.

*Corresponding author. † Equal contributors.

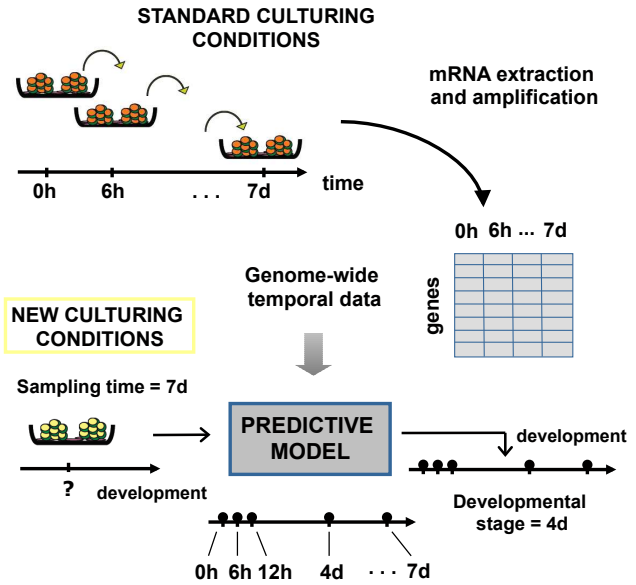



**Fig. 1.** Differentiation stage prediction models and their application. A prediction model is inferred from ESC genome-wide mRNA microarray data from experiments in standard culture conditions. In the illustration, the data are obtained along seven days of differentiation. The model can predict the cells developmental stage when the culture conditions are changed, that is, when the effects of a new chemical added to the culture are tested. Illustration shows the projection of the sample that, despite having been collected after seven days of differentiation, has a molecular identity of four days along the scale of standard differentiation.

## 2 METHODS

Let us consider a data set where the genome-wide expression profile has been observed for $n$ different samples along differentiation. Each sample is therefore represented with expression of thousands of genes. To infer a stage prediction model, we select the most informative genes and use data projection methods that combine the selected gene expression values and project sample profiles onto a one-dimensional ruler — a differentiation scale. Labels on the scale indicate the time at which the gene expression was measured and should, in standard culturing conditions, reflect the stages of cellular differentiation. Due to experimental noise and the variability of expression, samples from the same stage are not projected to the same point on the differentiation scale. To characterize the stage rather than the individual samples, these projections are fused into a single median position on the scale. The predictive model and its associated visualization through the differentiation scale can then predict the developmental stage of a sample coming from an experiment where different culturing conditions may have perturbed the differentiation process, and where the actual stage has yet to be determined (Figure 1).

In the following, we describe various data mining approaches we have considered for inference of predictive models. The transcriptome data typically includes measurements of a large number of genes for a small number of samples that were observed at different developmental stages. The inference starts with selection of most informative genes, that is, those that could be best used for the characterization of staging. Samples described with the selection of genes are then projected to the differentiation scale via unsupervised data mining methods or by additionally considering the stage information from the training data. We also introduce a variant of leave-two-out testing and concordance scoring to test and compare various model inference approaches.

### 2.1 Gene subset selection from time-sequence expression data

Our aim here is to reduce the computational costs of the inference and exclude from further analysis genes whose expression is either too constant or too irregular across the observed set of stages. A large variety of gene subset selection methods have been proposed for case-control studies (Cui and Churchill, 2003), but there are considerably fewer approaches for studies of expression dynamics. The distinguishing feature of such data sets is a low number of replicates observed at each of the developmental stages, which invalidates the utility of several standard statistical approaches, like those based on the analysis of variance (Park *et al.*, 2003; Nueda *et al.*, 2007; Cui and Churchill, 2003).

Perhaps the simplest approach, when the number of replicates per stage is low, is to rely on the expression fold change computed separately for each of the time points between two conditions (Gentile *et al.*, 2003; Peart *et al.*, 2005). These studies consider a gene to be differentially expressed during the time course if the measured fold change exceeds a selected threshold for at least one time point.

A powerful method using a statistical test was later proposed by Di Camillo *et al.* (2007). For each gene, it compares the area of the region bounded by the temporal expression profiles of treated and control cultures with a threshold based on a model of the experimental error. The method considers a gene to be differentially expressed if the observed area exceeds an estimated threshold. In order to remove the systematic bias introduced by experiments, the data must be normalized. In all of our experiments we used quantile normalization described by Bolstad *et al.* (2003).

In our study we have experimented with two different gene subset selection approaches, one that considers time (staging) and the other that ignores the series of events and treats stages as separate, unrelated experiments:

- AREA: As proposed by Di Camillo *et al.*, the area between the expression profile of a gene and its control profile was computed. The control profile was constant, equal to the expression at the first time point.
- FC: For each gene and each time point the fold change with respect to the control condition is computed. Gene expression at the initial time point was used as a control. Gene's score is defined as the number of time-points where expression change is at least two-fold.

The two scoring methods were used to, respectively, select the 1,000 best-ranked genes whose expression was then considered for inference of predictive models. The performance of prediction methods drops when a smaller number of informative genes is considered (see Supplementary data).

As a knowlegde-based alternative to data-driven gene subset selection, we have also identified a small set of *differentiation markers*, whose transcriptional signature could be considered for the inference of predictive models. These were obtained from Mouse Genome Informatics (MGI) repository (Bult *et al.*, 2010) from which we have retrieved the genes associated to stem cell differentiation (see `http://www.biolab.si/supp/stagerank`). Most of the associations were inferred by MGI on the basis of the Gene Ontology (term "stem cell differentiation" or its sub-terms such as "stem cell development" or "stem cell maintenance"). Similar procedure was applied to extract the markers lists for human, this time using Gene Ontology directly. The markers lists include a number of well known key pluripotency factors in mammal cells such as Sox2, Pou5f1 and Nanog (Pan and Thomson, 2007). Our experimental data sets included from 20 to 60 marker genes, and in this part of the experiments, only these genes were used for inference of stage prediction models.

### 2.2 Inference of prediction models

Let us assume we have a training data set represented with a $n \times m$ matrix $X$, where expression of $m$ genes has been observed for $n$ different samples. Let each sample be labeled by the development stage at which the measurements were performed. These can be placed in a column vector $Y$ of size $n$.

Our typical data set would contain about 5 to 12 different development stages and around 10 to 50 samples (typically, 3 or fewer samples per stage). Each sample would typically be represented with expression of 5,000 to 25,000 genes, from which we select 1,000 most informative ones using gene subset selection methods. Our goal is to project the samples to a 1-dimensional space, that is, place each sample on a line. Two distinct families of modeling approaches, called supervised and unsupervised methods, may address this problem with and without the knowledge of development stage ($Y$), respectively.

**Unsupervised** methods reduce the dimensionality of the data without considering the sample labels (stages). *Principal component analysis* (PCA) is probably one of the best known methods from this category. It linearly projects samples into a low-dimensional space that explains the highest degree of variance in the original data. We used the first principal component to project the samples into a single dimension. The projection of a sample $x$ is computed as $p(x) = xV$, where the column vector $V$ contains the first eigenvector of the covariance matrix $X^T X$, for mean-centered training data $X$.

As an alternative unsupervised method, we have also considered Pathrecon (Magwene *et al.*, 2003). Pathrecon starts by constructing a complete weighted graph with samples as nodes and their expression profile-based distances as edge weights. Then it finds a minimum spanning tree that connects all the nodes and includes edges such that a sum of their weights is minimized. The longest path in the tree is called a diameter path. Similarly to PCA's principal direction, diameter path orders the samples (nodes), but unlike PCA – and to the possible advantage of Pathrecon – the ordering is not constrained to a linear projection. Samples contained in the branches off the diameter path are assigned the same ordering index as the diameter path element to which they connect. If long off-diameter branches exist, a data structure called PQ-tree is used to summarize the uncertainties of path variations. Pathrecon traverses the PQ-tree to find candidate orderings, and ranks them by the distance of the path they describe.

Another approach we have considered is Minimum Curvilinear Embedding (MCE), a nonlinear dimension reduction method proposed by Cannistraci *et al.* (2010). The dimension reduction is performed by embedding high-dimensional data points into a lower dimensional space using the multidimensional scaling (MDS) algorithm. The data distances for MDS are computed as the traversal distances over the minimum spanning tree, which is constructed from Euclidean or Pearson correlation-based distances.

**Supervised** dimension reduction techniques use additional information on sample labels ($Y$). Since we aim at single dimension projections, we can represent successive labels $Y$ with their real-valued variants and use any regression algorithm. The inferred regression model maps a transcription profile to a real-value, in this way projecting the sample to an already defined differentiation scale. We aim to find the projection that best separates the different development stages. Since we have many more genes than samples ($n \ll m$), it is very easy to obtain a good separation and overfit the training data. *Partial least squares* (PLS) regression is known to work well even in such situations (Hőskuldsson, 1988), and does not overfit due to the high bias (linearity) in the description of the model. PLS is closely related to PCA and hence provides a good supervised counterpart. The particular variant of PLS used in our work is commonly referred to as PLS1 (Rosipal and Krämer, 2006), since the outcome matrix has only one column. In short, PLS1 first obtains a low-dimensional representation of $X$ by projecting it to a small number of latent variables. Then it models $Y$ as a linear combination of the latent variables. Computing the prediction for a new sample is done the same way: the values of the latent variables are calculated first and their weighted sum gives us the predicted result. For real-valued labels of development stages, we tested two different approaches. In the first we used the time (in hours) at which the samples were measured. In the second approach the consecutive developmental stages were represented with indices (e.g. 0,1,2,...).

Specialized methods have been proposed for learning ranking functions (Fürnkranz and Hüllermeier, 2010; Joachims, 2002; Cohen *et al.*, 1999). Ranking SVM (Joachims, 2002) is one of the earliest examples and is still considered a state-of-the-art approach and widely used as a benchmark for other rank learning methods. It tries to find a ranking function that maximizes Kendall's $\tau$ or, equivalently, minimizes the number of discordant pairs. Although this is NP-hard it can be approximated with a slight modification of the optimization problem. It turns out that the result is equivalent to considering the ranking problem as a binary classification problem on pairs of samples. In this context each pair is represented as a difference vector and plays the role of a single example in the standard classification SVM. For our experiments we used the freely available implementation $\mathrm{SVM}^{\mathrm{rank}}$, the ranking counterpart of the well known SVM package $\mathrm{SVM}^{\mathrm{light}}$ (`http://svmlight.joachims.org/`).

## 2.3 Evaluation and model scoring

We have experimentally compared various techniques for construction of stage prediction models. We used a number of gene expression data sets for testing, and performed evaluation either within the same data set (internal validation), or developed a model on one and tested the predictions on a different data set (external validation).

For the internal validation we use a variant of the *leave-pair-out* (LPO) approach (Pahikkala *et al.*, 2008). The procedure chooses two developmental stages, removes all samples from these two stages from the data set thus obtaining the training data, performs gene selection on the training data and then infers a prediction model. Finally, it tests the model on the samples from the two stages that were left out. We repeat this procedure for all different stage pairs. Notice that our implementation of leave-pair-out differs from the standard one which would leave out the samples regardless of their stages. Our concern here was that while retaining several samples from the specific stage in the training data, prediction of samples from that stage in the test set would have an advantage due to the potentially high similarity of same-stage samples. Staged leave-pair-out is thus more stringent, and in this respect even pessimistic: in real applications the models may be presented with samples that do belong to the stage that was also described in the training data.

For external validation, the prediction model is first developed on a selected training set. The model is then used to order the samples in the second (external) test set, where the quality of predictions are scored accordingly.

Pathrecon and MCE establish the ordering, but do not explicitly provide the model for staging. To enable stage prediction, we have included both the training and test samples in the input data, and determined the staging for the test samples from the obtained ordering.

For scoring of quality of predictions we use the concordance score $C$, a generalization of the the area under the receiver operating characteristic curve (*area under the curve* in short, AUC), a standard model discrimination measure. $C$ score is equal to the proportion of sample pairs for which the ranking by a prediction model corresponds to the true ranking, which is the same as the interpretation of $AUC$ (Hanley and McNeil, 1982). This interpretation also provides us with the means for its computation in the case of our particular testing procedures. We can check the ranking of two samples only if they come from two different stages of development. The sample pair is then ordered correctly if the order of projections corresponds to the order of the original stages. Formally, the score is computed as

$$C = \frac{\sum_{x \in T_i, y \in T_j, i<j} \delta(p(x) < p(y))}{\sum_{i<j} |T_i| \times |T_j|}, \qquad (1)$$

where $T_i$ is the set of samples from time point $i$, $p(x)$ a 1-dimensional projection of the sample $x$, $|T_i|$ the size of $T_i$ and $\delta(cond)$ equal to 0 or 1 if $cond$ is $False$ or $True$, respectively.

Computing the $C$ score with Eq. 1 works well in combination with the LPO cross-validation since we only need to check results for a pair of samples at a time. We also get good, unbiased score estimates even when evaluation is done on smaller data sets (Airola *et al.*, 2009).

## 3 EXPERIMENTAL ANALYSIS

We have evaluated different combinations of three gene selection methods (data-driven, FC and AREA, and marker-based, Markers) and four modeling techniques (PCA, PLS, SVMRank, MCE). In addition, PLS used either the actual time values ("time" in the name of the method) or stage indices. Pathrecon was run with authors' own implementation (Magwene *et al.*, 2003) on entire data sets. We here report MCE with Euclidean distance as it performed better than correlation-based distance. Also, only PCA is reported in combination with marker-genes. Entire set of experimental results with all possible combinations of gene selection, modeling methods, and distances for MCE is provided in the Supplementary data. The methods were tested on data sets from Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo`).

### 3.1 Data

Several data sets deposited in GEO focus on complex biological processes evolving over time, such as disease progression, development and cell differentiation, and thus provide the gene expression time series which could benefit from the construction of development stage prediction models. From a larger collection of such data sets, we have considered only those with at least six time points (stages) and with at least three samples for each stage. We foresee that one of the most promising applications of our work is the prediction of developmental potency of ESCs. We were thus more interested in experiments on cell development than in studies in which the behavior of cells under different treatments or in different disease states is analyzed over time. For this reason, we did not consider data on case-control studies but have analyzed only time series experiments of different organisms.

Ten data sets from different species meet these criteria and were chosen for our evaluation (GDS2666, GDS2667, GDS2668, GDS2669, GDS2671, GDS2672, GDS586, GDS587, GDS2431, GDS2688). Most of these data sets study the differentiation of mouse ESCs. In particular, the first six have been collected by Hailesellasse Sene *et al.* (2007) to study 11 stages of differentiation into embryoid bodies for three biologically equivalent but genetically distinct mouse ESC lines (R1, J1 and V6.5). The compatibility in the type of experiment and microarray data of these six data sets allowed to carry out external validation, that is, assess the predictive models trained from one data set through the quality of predictions on another data set. Data sets GDS586 and GDS587 analyze gene expression in a 12-day time course of mouse differentiating myoblasts. The last two data sets included in our analysis contain human and rat data, respectively. In GDS2431 the authors monitor gene expression in developing human erythroid progenitors, while for data set GDS2688 they analyze the temporal response of skeletal muscles to corticosteroid exposure in rats for up to 7 days. Although the aim of the latter study was different from the other cell differentiation data sets, it also had enough time points and replicates and we decided to include it for comparison.

In a separate experiment, we used the data sets from the study by Aiba *et al.* (2009) (GSE11523). From their collection of samples, we selected four cell lines (N, Z, G, F) that included at least three stages of cellular differentiation into specific germ layer types. Three of these cell lines (N, Z and G) consisted of ESCs differentiating into primitive and neural ectoderm, trophoblast, and primitive endoderm, respectively. With the F cell line the researchers analyzed a different stem cell type, the embryonic carcinoma stem cells, while undergoing a differentiation into primitive endoderm. These authors have already shown that the samples from the same cell lines project nicely and consistently in three dimensional space, and that the trajectory could qualitatively indicate the developmental potency of mouse ESCs. We adopt their data in order to quantitatively and systematically assess the quality of such predictions. In their original study Aiba *et al.* also show that the principal component-projected trajectories diverge for different cell lines when visualized in three dimensions. We were still interested if, despite this divergence, the predictive models developed on one cell-line maintain their stage prediction quality when predicting on the data from other cell lines.

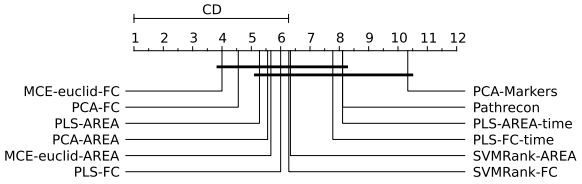



**Fig. 2.** Critical difference graph for method ranks from Table 1. Critical difference (CD) indicates the difference in ranks that would separate two significantly different approaches ($p < 0.05$).

### 3.2 Assessment of predictive accuracy

We report the $C$ scores for different internal and external validations. Table 1 summarizes results of the internal validation on the Gene Expression Omnibus data sets. For each data set, we ranked the methods according to the achieved $C$ score, and then report the average rank. Data set GDS2688 was not included in these averages as Pathrecon's score for it could not be computed in a reasonable amount of time (one day). The score for GDS2688 when using known markers instead of gene selection is not given, since the stem cell differentiation markers are not relevant for the process studied in this data set. The statistical analysis (Demšar, 2006) of rank differences is presented in Figure 2.

Results of external validation for six selected data sets are summarized in Table 2. Due to the insignificant differences in performance of different best-ranked methods we have here only used PCA for development of predictive models.

Similar analysis was also performed on data sets from Aiba *et al.* (2009). Again, for the internal validation, there were no significant differences in performance of various best-ranked methods considered ($p < 0.05$). For brevity, Table 3 compares only the scores for the six best-ranked methods from Table 1. Results of external validation are given in Table 4. As before, only the performance of the PCA-inferred model is reported.

In order to limit the uncontrolled sources of variability due to different microarrays platforms and experimental protocols the data sets used for external validation should refer to the same experimental setting. For this reason, we have kept the two sets of experiments (Tables 2 and 4) separated. However, the complete results for all possible pairs of data sets are available in the Supplementary data.

**Table 1.** C scores of leave-pair-out internal validation on ten different data sets from Gene Expression Omnibus. For each data set the methods are ranked according to data set-specific C score. Methods' average ranks and average C scores are also reported.

| | GDS2431 | GDS2666 | GDS2667 | GDS2668 | GDS2669 | GDS2671 | GDS2672 | GDS2688 | GDS586 | GDS587 | $\overline{C}$ | $\overline{rank}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCE-euclid-FC | **0.993** | 0.972 | **0.964** | 0.897 | 0.895 | **0.964** | **0.939** | **0.750** | 0.853 | 0.825 | 0.922 | 4.000 |
| PCA-FC | 0.874 | **0.974** | 0.899 | **0.931** | **0.909** | 0.822 | 0.794 | 0.732 | 0.948 | **0.942** | 0.899 | 4.556 |
| PLS-AREA | 0.867 | 0.945 | 0.913 | 0.923 | 0.903 | 0.909 | 0.812 | 0.581 | 0.944 | 0.884 | 0.900 | 5.278 |
| PCA-AREA | 0.896 | 0.952 | 0.921 | 0.929 | 0.889 | 0.824 | 0.798 | 0.738 | 0.944 | 0.937 | 0.899 | 5.556 |
| MCE-euclid-AREA | 0.941 | 0.966 | 0.941 | 0.901 | 0.877 | 0.911 | 0.828 | 0.728 | 0.817 | 0.820 | 0.889 | 5.667 |
| PLS-FC | 0.859 | 0.962 | 0.883 | 0.905 | **0.909** | 0.911 | 0.764 | 0.588 | 0.948 | 0.857 | 0.889 | 6.000 |
| SVMRank-FC | 0.844 | 0.915 | 0.907 | 0.889 | 0.883 | 0.893 | 0.897 | 0.551 | **0.972** | 0.857 | 0.895 | 6.278 |
| SVMRank-AREA | 0.859 | 0.883 | 0.881 | 0.893 | 0.905 | 0.859 | 0.913 | 0.542 | 0.964 | 0.862 | 0.891 | 6.333 |
| PLS-FC-time | 0.859 | 0.966 | 0.786 | 0.766 | 0.903 | 0.871 | 0.782 | 0.423 | 0.960 | 0.815 | 0.856 | 7.778 |
| Pathrecon | 0.956 | 0.840 | 0.887 | 0.859 | 0.812 | 0.919 | 0.784 | N/A | 0.897 | 0.804 | 0.862 | 8.111 |
| PLS-AREA-time | 0.867 | 0.952 | 0.766 | 0.760 | 0.798 | 0.863 | 0.842 | 0.392 | 0.952 | 0.841 | 0.849 | 8.111 |
| PCA-Markers | 0.600 | 0.911 | 0.869 | 0.877 | 0.842 | 0.887 | 0.776 | N/A | 0.730 | 0.519 | 0.779 | 10.333 |

**Table 2.** C scores for the PCA-AREA inferred models developed on a training set (row label) and tested on an independent test set (column label). Labels in superscripts of the scores denote the relationship between the two data sets: $^a$ same cell line, different platform; $^b$ different cell line, same platform; $^c$ different cell line, different platform.

| | GDS2666 | GDS2667 | GDS2668 | GDS2669 | GDS2671 | GDS2672 |
|---|---|---|---|---|---|---|
| GDS2666 | / | $0.915^a$ | $0.939^b$ | $0.901^b$ | $0.840^c$ | $0.818^c$ |
| GDS2667 | $0.947^a$ | / | $0.949^c$ | $0.909^b$ | $0.869^c$ | $0.857^b$ |
| GDS2668 | $0.980^b$ | $0.891^c$ | / | $0.893^a$ | $0.770^b$ | $0.804^c$ |
| GDS2669 | $0.941^c$ | $0.954^b$ | $0.941^a$ | / | $0.828^c$ | $0.830^b$ |
| GDS2671 | $0.958^b$ | $0.921^c$ | $0.954^b$ | $0.875^c$ | / | $0.711^a$ |
| GDS2672 | $0.941^c$ | $0.960^b$ | $0.935^c$ | $0.909^b$ | $0.840^a$ | / |

**Table 3.** LPO-validation C scores and comparison of four different modeling methods on data sets from Aiba *et al*. Methods' average C score across different data sets and average rank are reported.

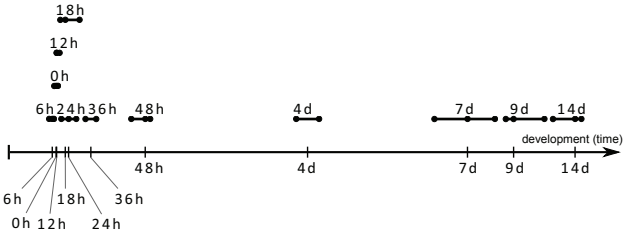
| | F | G | N | Z | $\overline{C}$ | $\overline{rank}$ |
|---|---|---|---|---|---|---|
| PLS-FC | 0.883 | **1.000** | 0.905 | **0.983** | 0.943 | 2.875 |
| PCA-AREA | 0.883 | **1.000** | **0.917** | 0.917 | 0.929 | 2.875 |
| PLS-AREA | 0.950 | 0.933 | 0.905 | 0.950 | 0.935 | 3.125 |
| MCE-euclid-AREA | **0.983** | 0.700 | 0.905 | 0.583 | 0.793 | 4.000 |
| MCE-euclid-FC | **0.983** | 0.733 | 0.905 | 0.533 | 0.789 | 4.000 |
| PCA-FC | 0.883 | 0.867 | 0.857 | **0.983** | 0.898 | 4.125 |

**Table 4.** C scores of external validation for PCA-AREA models inferred on data sets from Aiba *et al*. Training sets (row labels) and test sets (column labels) represent different cell lines measured with the same experimental platform.

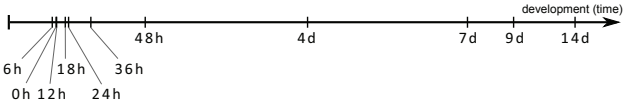
| | F | G | N | Z |
|---|---|---|---|---|
| F | / | 1.000 | 0.976 | 1.000 |
| G | 0.933 | / | 1.000 | 0.950 |
| N | 0.850 | 0.883 | / | 0.817 |
| Z | 0.767 | 0.750 | 0.988 | / |

### 3.3 Analysis of inferred differentiation scales

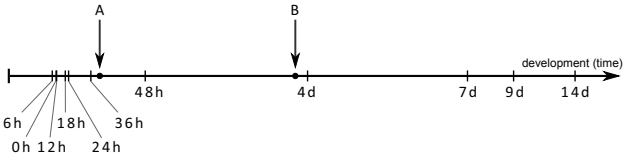
From the five modeling methods considered, PCA and MCE are the only ones that truly discover the relations between cell stages from the data, that is, constructs an informative differentiation scale. PLS and SVMRank are supervised and perhaps focus too much on optimizing their respective goals. For example, while expression profiles taken after 18 and 24 hours might be very similar, supervised algorithms will still try to separate the projections, because they know the samples come from different time-points. While Pathrecon is unsupervised, it only orders samples and does not provide a model for projection. MCE can be used for projection, but does not provide an explicit model for staging of new samples. That is why we here examine only PCA's differentiation scales. For all of the examined data sets, we found that the scales order the stages very




(a) Projection of the samples from the training data illustrates the construction of the predictive model and its associated differentiation scale. For visual clarity, projections are arbitrarily vertically dispersed. Samples observed at the same development stage are connected with a line.




(b) Differentiation scale for mouse embryonic stem cell differentiation.




(c) Prediction of the developmental stage for two samples (A and B).

**Fig. 3.** An example of the projection of samples in a single-dimensional plot demonstrating the construction of the differentiation scale (a), the obtained differentiation scale representing a predictive model (b) and prediction of developmental stages of new samples (c).

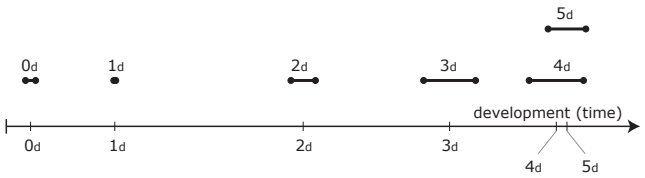



**Fig. 4.** Differentiation scale for *M. musculus* embryonal carcinoma stem cells differentiating into primitive endoderm (data set F).

well with only minor errors in the order of similar stages. For brevity we demonstrate this successful result on two selected data sets (Figures 3 and 4).

As a first example, let us illustrate the composition and utility of the differentiation scale and associated prediction model with the data from a study of the mouse R1 ESC line. The data included 11 different time points during 14 days of differentiation into embryoid bodies (EBs) (Hailesellasse Sene *et al.*, 2007). At each time point the data (GDS2667) contains measurements of over 18,000 genes in three different biological replications. The predictive model was inferred from the data comprising the entire set of 33 samples from which we have excluded two samples for testing purposes. The projections in Figure 3 were inferred using PCA-AREA on a subset of the 1,000 most informative genes. The Gene Ontology annotation of this group of selected genes highlighted the efficacy of the selection strategy, with a significant number of genes annotated to biological functions involved in cellular differentiation, such as developmental process (25% of genes), growth (17%), and apoptosis (8%). The time-ticks in the differentiation scale in Figure 3(b), which indicate the developmental stages of the cell, correspond to the median position of the projections of samples taken at the same stage of development. They are ordered as expected, except for one transposition of the very similar time points at 0 and 6 hours. We can also observe a wide gap around 4 days of development, most probably reflecting the specific time resolution used in the experiment, but also indicating that the cells undergo a substantial change in the time period between 48 hours and the 7th day.

To show the predictive capability of the model, we have left out two samples (A and B, Figure 3(c)) which were measured at 36 hours and 4 days, respectively. They were correctly projected, thus validating the model as a predictive tool.

For a second example we study a different type of pluripotent cells, the F9 embryonal carcinoma cell line during its differentiation into parietal endoderm for 5 days (Gene Expression Omnibus, GSE11523). The projection of this data set results in a perfectly monotonic scale of development. The last two stages do overlap, but this could be explained by the intrinsic variability in the speed of differentiation and in the composition of the three germ layers of each single cell line. This variation increases during differentiation and affects DNA microarray measurements.

### 3.4 Prediction of differentiation of induced pluripotent stem cells

To further test the utility of proposed models for stage prediction, we used transcription data from induced pluripotent stem cells (IPSCs), another type of pluripotent cells. These cells are obtained by the forced expression of four pluripotency factors in differentiated somatic cells and share with ESCs the same pluripotent potential (Takahashi and Yamanaka, 2006; Okita *et al.*, 2007). Given the great interest for their possible therapeutical use in the production of patient-specific cell lines that could be transplanted without rejection, many investigations have been carried out since their discovery. These studies have highlighted the diverse pluripotent status of different IPSC lines when created in separated laboratories or with slightly different protocols. Using the prediction model developed from embryonic R1 stem cells differentiation in vitro (GDS2666, (Hailesellasse Sene *et al.*, 2007)), we assessed the pluripotency status of an IPSC line (Zhao *et al.*, 2009) recently generated from mouse embryonic fibroblasts (MEF). GDS2666 was obtained with a microarray chip different from that of the IPSCs, so prior to the projection the data was scaled using global scale normalization (Yang *et al.*, 2002). Projection to the differentiation scale (Fig. 5) confirmed the pluripotency of the IPSCs, positioning the projection within the 0-36 h time interval. On the contrary, also confirming the utility of proposed prediction method, the projection of the differentiated MEF cells fell to the “differentiated” part of the scale, within the 7-9 day interval.

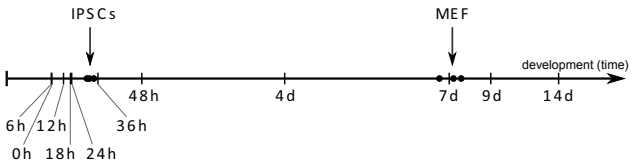



**Fig. 5.** Projection of IPSC samples and adult MEF samples on the differentiation scale of ESCs cultured in standard conditions (GDS2666). Three biological replicates for each sample were processed.

## 4 DISCUSSION

The predictive accuracy of inferred models is very high when they are applied to data from the same cell lines as used in the training set. The reasonable range of $C$ scores is from 0.5 (random predictor) to 1.0 (perfect prediction). The majority of $C$ scores for the described methods are close to 0.9, a very high score indicating an excellent quality of predictions. The only notable exception is data set GDS2688, where all methods achieved lower scores. PCA and MCE were here the only two methods that obtained reasonably good results. We can conclude that predicting the stage of development from transcriptional profiles is feasible and that the resulting prediction models can accurately predict developmental stages within a chosen cell line.

The results of external validation are also interesting. Aiba *et al.* observed that the trajectories obtained from cells of different cell lines diverged to a large extent. We therefore expected that the predictions of models developed on one cell line would fail when applied to data from another cell line. Results on our selection of data sets from GEO (Table 2) refute this expectation, and demonstrate that the tested predictive models can be applied across different cell lines. In addition, external validation on data sets from Aiba *et al.* (2009) (Table 4) is also qualitatively similar, showing that prediction across different cell lines is indeed feasible and may be highly accurate. The scores we have obtained are surprisingly high, with only four below 0.80 and 24 above 0.90 (out of 42). The results for all possible pairs of data sets shown in the Supplementary data confirm the high accuracy of predictions even for data coming from different studies. Poor predictions were obtained only for training and test data from different species.

Utility of stage prediction models across different cell lines was further confirmed in IPSCs and MEF experiments. Projection of related transcription profiles on a PCA-inferred differentiation scale highlighted the difference in pluripotency between the adult cells and the reprogrammed cell line.

Among the tested methods the differences in predictive quality were not statistically significant. PCA, MCE, PLS and SVMRank are all time-efficient and construct corresponding models for the data sets in our study within seconds. Pathrecon can be very slow with execution times of several hours or even days for data sets where construction of a PQ-tree and examination of all candidate orderings is required. At the present stage of evaluation, we thus prefer the principal component analysis because of its simplicity, explicit prediction model, and the added benefit of its informative differentiation scales (Figures 3 and 4). The staging is easy to interpret by biologists, and the visualization uncovers the dynamics of the changes with phenotypically different stages being placed farther apart on the differentiation scale. As its non-linear couterpart, MCE looks very promising and should be considered along PCA in further studies of this kind.

## 5 CONCLUSION

Developmental biology is in need of devices that would accurately assess the progression of cells through development, and predict the developmental stages of cells observed under different physiological conditions. We have proposed and investigated the utility of approaches that can make such predictions. Experiments show that the differentiation stage prediction models inferred from transcription profiles are feasible, have high accuracy, and that their results can be nicely mapped to simple, one-dimensional differentiation scales.

## ACKNOWLEDGEMENT

*Funding*: This work was supported by the Fondazione Cariplo Project: “Bioinformatics for Tissue Engineering: Creation of an International Research Group”, by the FIRB project “ITALBIONET”, and by the grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112). We thank Gad Shaulsky and Lucia Sacchi for useful comments and suggestions.

## REFERENCES

Aiba, K., Nedorezov, T., Piao, Y., Nishiyama, A., Matoba, R., Sharova, L. V., Sharov, A. A., Yamanaka, S., Niwa, H., and Ko, M. S. H. (2009). Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells. *DNA Res*, **16**(1), 73–80.

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2009). A comparison of auc estimators in small-sample studies. In *Proceedings of the Third International Workshop on Machine Learning in Systems Biology (MLSB’09)*, pages 15–23.

Bhattacharya, B., Puri, S., and Puri, R. K. (2009). A review of gene expression profiling of human embryonic stem cell lines and their differentiated progeny. *Current stem cell research & therapy*, **4**(2), 98–106.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–93.

Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A., Eppig, J. T., and the Mouse Genome Database Group (2010). The mouse genome database: enhancements and updates. *Nucleic Acids Research*, **38**(suppl 1), D586–D592.

Cannistraci, C. V., Ravasi, T., Montevecchi, F. M., Ideker, T., and Alessio, M. (2010). Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics*, **26**(18), i531–i539.

Cohen, W. W., Schapire, R. E., and Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, **10**(1), 243–270.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**(4), 210.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, **7**, 1–30.

Di Camillo, B., Toffolo, G., Nair, S. K., Greenlund, L. J., and Cobelli, C. (2007). Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics*, **8**(Suppl 1), S10.

Fürnkranz, J. and Hüllermeier, E., editors (2010). *Preference Learning*. Springer-Verlag.

Gentile, M., Latonen, L., and Laiho, M. (2003). Cell cycle arrest and apoptosis provoked by UV radiation-induced DNA damage are transcriptionally highly divergent responses. *Nucleic Acids Res*, **31**(16), 4779–90.

Hailesellasse Sene, K., Porter, C. J., Palidwor, G., Perez-Iratxeta, C., Muro, E. M., Campbell, P. A., Rudnicki, M. A., and Andrade-Navarro, M. A. (2007). Gene

function in early mouse embryonic stem cell differentiation. *BMC Genomics*, **8**, 85.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1), 29–36.

Hőskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, **2**(3), 211–228.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142.

Magwene, P. M., Lizardi, P., and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, **19**(7), 842–850.

Mata, J., Lyne, R., Burns, G., and Bähler, J. (2002). The transcriptional program of meiosis and sporulation in fission yeast. *Nat Genet*, **32**(1), 143–7.

Müller, F., Laurent, L., Kostka, D., Ulitsky, I., Williams, R., Lu, C., Park, I., Rao, M., Shamir, R., Schwartz, P., Schmidt, N., and Loring, J. (2008). Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**(7211), 401–405.

Neri, T., Merico, V., Fiordaliso, F., Salio, M., Rebuzzini, P., Sacchi, L., Bellazzi, R., Redi, C., Zuccotti, M., and Garagna, S. (2011). The differentiation of cardiomyocytes from mouse embryonic stem cells is altered by dioxin. *Toxicol Lett.*, **202**(3), 226–236.

Novershtern, N., Subramanian, A., Lawton, L., R.H., M., Haining, W., McConkey, M., Habib, N., Yosef, N., Chang, C., Shay, T., Frampton, G., Drake, A., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J., Liefeld, T., Smutko, J., Chen, J., Friedman, N., Young, R., Golub, T., Regev, A., and Ebert, B. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**(2), 296–309.

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C. J., Smilde, A. K., Talón, M., and Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics*, **23**(14), 1792–800.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, **448**(7151), 313–317.

Pahikkala, T., Airola, A., Boberg, J., and Salakoski, T. (2008). Exact and efficient leave-pair-out cross-validation for ranking RLS. In *Proceedings of the 2nd international and interdisciplinary conference on adaptive knowledge representation and reasoning (AKRR'08)*, pages 1–8.

Pan, G. and Thomson, J. A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res*, **17**(1), 42–49.

Park, T., Yi, S.-G., Lee, S., Lee, S. Y., Yoo, D.-H., Ahn, J.-I., and Lee, Y.-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, **19**(6), 694–703.

Peart, M. J., Smyth, G. K., van Laar, R. K., Bowtell, D. D., Richon, V. M., Marks, P. A., Holloway, A. J., and Johnstone, R. W. (2005). Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proc Natl Acad Sci U S A*, **102**(10), 3697–702.

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. a., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. a., Ideker, T., and Hayashizaki, Y. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**(5), 744–752.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51. Springer Berlin / Heidelberg.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**(4), 663–676.

Van Driessche, N., Shaw, C., Katoh, M., Morio, T., Sucgang, R., Ibarra, M., Kuwayama, H., Saito, T., Urushihara, H., Maeda, M., Takeuchi, I., Ochiai, H., Eaton, W., Tollett, J., Halter, J., Kuspa, A., Tanaka, Y., and Shaulsky, G. (2002). A transcriptional profile of multicellular development in dictyostelium discoideum. *Development*, **129**(7), 1543–52.

Wagner, R. A., Tabibiazar, R., Liao, A., and Quertermous, T. (2005). Genome-wide expression dynamics during mouse embryonic development reveal similarities to drosophila development. *Dev Biol*, **288**(2), 595–611.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4), e15.

Zhao, X., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C., Ma, Q., Wang, L., Zeng, F., and Zhou, Q. (2009). iPS cells produce viable mice through tetraploid complementation. *Nature*, **461**(7260), 86–90.