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[2,5] and Blaz Zupan[1,2,6,*]

[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, [2]Centre for Tissue Engineering, University of Pavia, Pavia, [3]Dipartimento di Biologia Animale, Laboratorio di Biologia dello Sviluppo, University of Pavia, Pavia, [4]Sezione di Istologia ed Embriologia, Dipartimento di Medicina Sperimentale, University of Parma, Parma, [5]Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia, Italy and [6]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

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 cell's 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 datasets 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 and accompanying datasets 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. In the early stages of mammal differentiation, the inner cell mass undergoes gastrulation. In this process, three specific germ layers—mesoderm,

---

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

endoderm and ectoderm—are formed. These layers ultimately give rise to specific tissues and organs. 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 favor of one that defines a more differentiated type of cellular identity. Novershtern *et al.* (2011) demonstrated 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 (Bhattacharya *et al.*, 2009; Cannistraci *et al.*, 2010; Mata *et al.*, 2002; Neri *et al.*, 2011; Ravasi *et al.*, 2010; Van Driessche *et al.*, 2002; Wagner *et al.*, 2005). 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 1D projection. In this article, 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. The differentiation scale depicts how far the cells have departed from the origin, the embryonic pluripotent state. The proposed methods do not explicitly distinguish between any direction of differentiation in terms of composition of the different germ layers. Yet, if desired, these can be considered by developing the scale from direction-specific data.

Projection of differentiation landmarks on the differentiation scale may also 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 datasets from mouse, rat and human.

2 METHODS

Let us consider a dataset 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 1D 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 (Fig. 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 Formal definitions

Let us assume we have a training dataset 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 y_i at which the measurements were performed. These can be placed in a column vector Y of size n. Our typical dataset would contain about 5–12 different development stages and around 10–50 samples (typically, 3 or fewer samples per stage). Each sample would typically be represented with expression of 5000–25 000 genes, from which we select 1000 most informative ones using gene subset selection methods. Our goal is to project the samples to a 1D space, that is, assign a real number $p(x_i)$ to sample i based on its expression profile x_i (a row from X). The ordering of the samples by their projections should correspond to the order of their differentiation, i.e.:

$$p(x_i) > p(x_j) \Leftrightarrow y_i > y_j \qquad (1)$$

Different approaches may be used to obtain the projections $p(x_i)$. Model-based approaches infer projection models in the form of a function $p: \mathbb{R}^m \to \mathbb{R}$ that maps an expression profile to a real number. A different set of techniques do not infer explicit models, but instead directly compute either projections or ranks of the presented samples. The ordering constraint from Equation (1) may not be achieved for all pairs of samples i, j, so we define a scoring function to assess the quality of different inference methods.

2.2 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
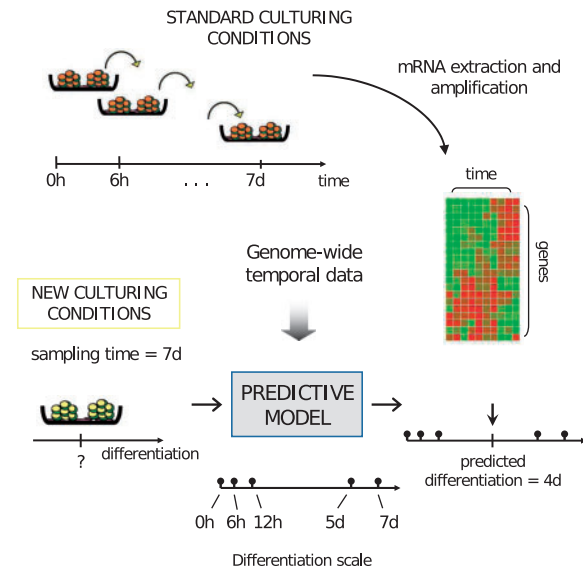

Fig. 1. Differentiation stage prediction models and their application. A prediction model is inferred from ESC genome-wide mRNA microarray data from experiments in standard culture conditions. In the illustration, the data are obtained during 7 days of differentiation. The model can predict the cell's stage in terms of how far the cells have departed from their embryonic pluripotent state. Such predictions may be useful when the culture conditions are changed, that is, when the effects of a new chemical added to the culture are tested. For example, despite having been collected after 7 days of differentiation, the cell culture from the figure has an estimated molecular identity of 4 days along the scale of standard differentiation.

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 when a low number of replicates at each developmental stage is observed.

After applying quantile normalization (Bolstad *et al.*, 2003), 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.* (2007), the area of the region bounded by the temporal 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 was computed (Gentile *et al.*, 2003; Peart *et al.*, 2005). 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 2-fold.

The two scoring methods were used to, respectively, select the 1000 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 (Supplementary Material).

As a knowledge-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 (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 subterms 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 datasets 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.

2.3 Inference of prediction models

To address the problem of finding suitable projections with and without the knowledge of development stage (Y), respectively, we have investigated the utility of two distinct families of modeling approaches, called supervised and unsupervised methods.

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 non-linear 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. MCE is a parameter-free projection algorithm and was shown to be particularly effective in discriminating the classes in small-*n* large-*m* datasets by using only the first dimension (Cannistraci *et al.*, 2010).

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 (Höskuldsson, 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 (Cohen *et al.*, 1999; Fürnkranz and Hüllermeier, 2010; Joachims, 2002). 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 τ 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 SVM^rank, the ranking counterpart of the well-known SVM package SVM^light (http://svmlight.joachims.org/).

2.4 Evaluation and model scoring

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

$$C = \frac{\sum\limits_{x \in T_i, y \in T_j, i < j} \delta(p(x) < p(y))}{\sum\limits_{i < j} |T_i| \times |T_j|}, \tag{2}$$

where T_i is the set of samples from time point i, $p(x)$ a 1D 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 Equation (2) 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 datasets (Airola *et al.*, 2009).

Predictive accuracies are estimated on different datasets. To compare the performance of multiple methods on multiple datasets, we use the Friedman test as described in Demšar (2006). The methods are ranked for each dataset separately and their average ranks compared. We use Nemenyi test as a *post hoc* test to compute the required difference in ranks of two methods for their performance to be considered significantly different. A critical difference graph (Demšar, 2006) is used for a visual depiction of the results.

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 datasets. 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 Material. The methods were tested on datasets from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo).

3.1 Data

Several datasets 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 datasets, 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 datasets from different species met these criteria and were chosen for our evaluation (GDS2666, GDS2667, GDS2668, GDS2669, GDS2671, GDS2672, GDS586, GDS587, GDS2431, GDS2688). Most of these datasets 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 datasets allowed to carry out external validation, that is, assess the predictive models trained from one dataset through the quality of predictions on another dataset. Datasets GDS586 and GDS587 analyze gene expression in a 12-day time course of mouse differentiating myoblasts. The last two datasets included in

our analysis contain human and rat data, respectively. In GDS2431, the authors monitor gene expression in developing human erythroid progenitors, while for dataset 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 datasets, it also had enough time points and replicates and we decided to include it for comparison.

In a separate experiment, we used the datasets 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 3D 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.

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 datasets. For each dataset, we ranked the methods according to the achieved C score, and then report the average rank. Dataset 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 dataset. The statistical analysis of the methods' performance ranks is summarized in Figure 2.

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

Similar analysis was also performed on datasets 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 datasets 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 datasets are available in the Supplementary Material.

Table 1. *C* scores of LPO internal validation on 10 different datasets from Gene Expression Omnibus

	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

For each dataset, the methods are ranked according to dataset-specific *C* score. Methods' average ranks and average *C* scores are also reported. Best score for each dataset is printed in bold. Scores for entries marked with N/A are not available; see Section 3.2 for explanation.
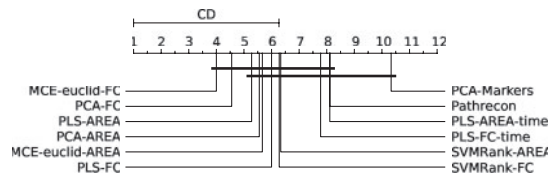

Fig. 2. Critical difference graph for method ranks from Table 1. Critical difference (CD) indicates the difference in ranks that would separate two significantly different approaches ($P < 0.05$).

3.3 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. For all of the examined datasets, we found that the scales order the stages very well with only minor errors in the order of similar stages. For brevity, we demonstrate the successful result of PCA's differentiation scales on two selected datasets (Figs 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 (GDS2666) 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 1000 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 3b, 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 h. We can

also observe a wide gap ∼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 h and the 7th day.

To show the predictive capability of the model, we have left out two samples (A and B, Figure 3c) that were measured at 36 h 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 dataset 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 (Okita *et al.*, 2007; Takahashi and Yamanaka, 2006). Given the great interest for their possible therapeutic 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 (MEFs). 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

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)

	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]	/

Labels in superscripts of the scores denote the relationship between the two datasets:
[a]Same cell line, different platform.
[b]Different cell line, same platform.
[c]Different cell line, different platform.

Table 3. LPO-validation *C* scores and comparison of four different modeling methods on datasets from Aiba *et al*.

	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

Methods' average *C* score across different datasets and average rank are reported. Best score for each data set is printed in bold.

Table 4. *C* scores of external validation for PCA-AREA models inferred on datasets from Aiba *et al*.

	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	/

Training sets (row labels) and test sets (column labels) represent different cell lines measured with the same experimental platform.

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.

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 (Tables 1 and 3). 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 dataset GDS2688, where all methods achieved lower scores. This is not surprising as GDS2688 is substantially different from the other datasets both in the type of cells and the processes
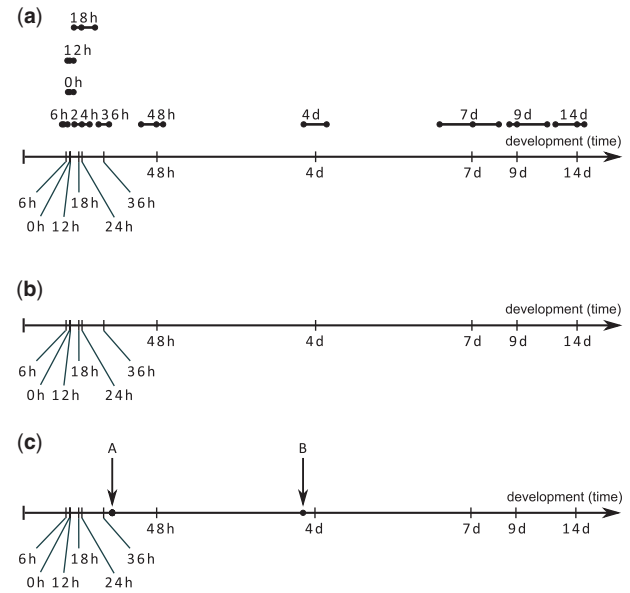

Fig. 3. An example of the projection of samples in a single-dimensional plot demonstrating the construction of the differentiation scale (**a**), the obtained differentiation scale representing a predictive model (**b**) and prediction of developmental stages of new samples (**c**).(a) Projection of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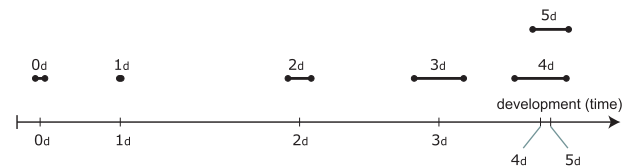

Fig. 4. Differentiation scale for *Mus musculus* embryonal carcinoma stem cells differentiating into primitive endoderm (dataset F).
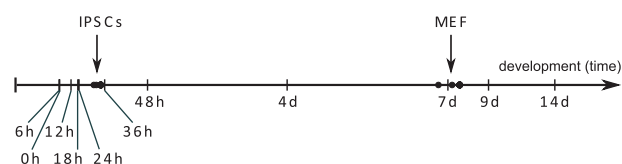
Fig. 5. Projection of IPSC samples and adult MEF samples on the differentiation scale of ESCs cultured in standard conditions (GDS2666). Three biological replicates for each sample were processed.

observed. 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 datasets 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 datasets 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 <0.80 and 24 >0.90 (out of 42). The results for all possible pairs of datasets shown in the Supplementary Material 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 datasets in our study within seconds. Pathrecon can be very slow with execution times of several hours or even days for datasets where construction of a PQ-tree and examination of all candidate orderings is required. PLS and SVMRank are supervised and may focus too much on optimizing their respective goals. For example, while expression profiles taken after 18 and 24 h 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.

At the present stage of evaluation, we thus prefer the PCA because of its simplicity, explicit prediction model and the added benefit of its informative differentiation scales (Figs 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 counterpart, MCE looks very promising and should be considered along PCA in further studies of this kind.

We have noticed some fluctuations of performance of several methods on different datasets. The differences could be arbitrary, or related to specific types of patterns hidden in the data that a particular method can or cannot detect. This issue could be further studied once more datasets of the kind examined in this article will become available.

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, 1D differentiation scales.

REFERENCES

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.

Airola,A. *et al.* (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)*, Jozef Stefan Institute, Ljubljana, Slovenia, pp. 15–23.

Bhattacharya,B. *et al.* (2009) A review of gene expression profiling of human embryonic stem cell lines and their differentiated progeny. *Curr. Stem Cell Res. Therapy*, **4**, 98–106.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bult,C.J. *et al.*; Mouse Genome Database Group (2010). The mouse genome database: enhancements and updates. *Nucleic Acids Res.*, **38**(Suppl. 1), D586–D592.

Cannistraci,C.V. *et al.* (2010) Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics*, **26**, i531–i539.

Cohen,W.W. *et al.* (1999) Learning to order things. *J. Art. Intell. Res.*, **10**, 243–270.

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

Demšar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

Di Camillo,B. *et al.* (2007) Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics*, **8**(Suppl. 1), S10.

Fürnkranz,J. and Hüllermeier,E. (eds) (2010). *Preference Learning*. Springer, Heidelberg.

Gentile,M. *et al.* (2003) Cell cycle arrest and apoptosis provoked by UV radiation-induced DNA damage are transcriptionally highly divergent responses. *Nucleic Acids Res.*, **31**, 4779–4790.

Hailesellasse Sene,K. *et al.* (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**, 29–36.

Hőskuldsson,A. (1988) PLS regression methods. *J. Chemomet.*, **2**, 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*, ACM New York, NY, pp. 133–142.

Magwene,P.M. *et al.* (2003) Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, **19**, 842–850.

Mata,J. *et al.* (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nat. Genet.*, **32**, 143–147.

Müller,F. *et al.* (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.

Neri,T. *et al.* (2011) The differentiation of cardiomyocytes from mouse embryonic stem cells is altered by dioxin. *Toxicol. Lett.*, **202**, 226–236.

Novershtern,N. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.

Okita,K. *et al.* (2007) Generation of germline-competent induced pluripotent stem cells. *Nature*, **448**, 313–317.

Pahikkala,T. *et al.* (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)*, Helsinki University of Technology, Espoo, Finland, pp. 1–8.

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

Peart,M.J. *et al.* (2005) Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proc. Natl Acad. Sci. USA*, **102**, 3697–3702.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Rosipal,R. and Krämer,N. (2006) Overview and recent advances in partial least squares. In Saunders,C. *et al.* (eds) *Subspace, Latent Structure and Feature Selection*, Vol. 3940 of *Lecture Notes in Computer Science.* Springer, Berlin/Heidelberg, pp. 34–51.

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

Van Driessche,N. *et al.* (2002) A transcriptional profile of multicellular development in dictyostelium discoideum. *Development*, **129**, 1543–1552.

Wagner,R.A. *et al.* (2005) Genome-wide expression dynamics during mouse embryonic development reveal similarities to drosophila development. *Dev. Biol.*, **288**, 595–611.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zhao,X. *et al.* (2009) iPS cells produce viable mice through tetraploid complementation. *Nature*, **461**, 86–90.