Results:

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

Lan Zagar1,†, Francesca Mulas2,†, Silvia Garagna3, Maurizio Zuccotti4,
Riccardo Bellazzi2,5 and Blaz Zupan1,2,6,*

1Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, 2Centre for Tissue
Engineering, University of Pavia, Pavia, 3Dipartimento di Biologia Animale, Laboratorio di Biologia dello Sviluppo,
University of Pavia, Pavia, 4Sezione di Istologia ed Embriologia, Dipartimento di Medicina Sperimentale, University of
Parma, Parma, 5Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia, Italy and 6Department 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.

Received on March 8, 2011; revised on July 7, 2011; accepted on
July 11, 2011

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
gastulation. In this process, three specific germ layers—mesoderm,
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 (Muller 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.

2546 © The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
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 fit a model-based projection 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.,

\[
p(x_i) > p(x_j) \implies y_i > y_j
\]

(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 \( p: \mathbb{R}^m \rightarrow \mathbb{R} \) 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
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 experiments 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 \( s \) is computed as \( p(s) = m - X \cdot C \), 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 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 the 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, it makes use of 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- to large-n datasets by using only the first dimension (Cannistraci et al., 2010).

In contrast to the previous methods, 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 (Hoskuldsson, 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; Flammknecht 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 \( \tau \) or, equivalently, minimizes the number of discordant samples. 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^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 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

\[
C = \frac{\sum_{i \lt j \in T_1 \times T_2} \delta(p(x) < p(y)) - \sum_{i \in T_1} |T_1| \times |T_2|}{|T_1| \times |T_2|} \tag{2}
\]
Stage prediction of stem cell differentiation

3 EXPERIMENTAL ANALYSIS

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MCE-euclid-FC</th>
<th>PCA-FC</th>
<th>PLS-AREA</th>
<th>SVMRank-AREA</th>
<th>SVMRank-FC</th>
<th>PLS-FC-time</th>
<th>Pathrecon</th>
<th>PLS-AREA-time</th>
<th>PCA-Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDS2431</td>
<td>0.993</td>
<td>0.875</td>
<td>0.964</td>
<td>0.897</td>
<td>0.905</td>
<td>0.990</td>
<td>0.762</td>
<td>0.885</td>
<td>0.915</td>
</tr>
<tr>
<td>GDS2666</td>
<td>0.964</td>
<td>0.952</td>
<td>0.942</td>
<td>0.952</td>
<td>0.950</td>
<td>0.968</td>
<td>0.950</td>
<td>0.969</td>
<td>0.956</td>
</tr>
<tr>
<td>GDS2667</td>
<td>0.897</td>
<td>0.903</td>
<td>0.905</td>
<td>0.909</td>
<td>0.903</td>
<td>0.910</td>
<td>0.911</td>
<td>0.908</td>
<td>0.905</td>
</tr>
<tr>
<td>GDS2686</td>
<td>0.785</td>
<td>0.822</td>
<td>0.809</td>
<td>0.824</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
</tr>
<tr>
<td>GDS2687</td>
<td>0.750</td>
<td>0.822</td>
<td>0.809</td>
<td>0.824</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
</tr>
<tr>
<td>GDS2671</td>
<td>0.750</td>
<td>0.822</td>
<td>0.809</td>
<td>0.824</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
<td>0.828</td>
</tr>
<tr>
<td>GDS2672</td>
<td>0.964</td>
<td>0.903</td>
<td>0.905</td>
<td>0.910</td>
<td>0.913</td>
<td>0.913</td>
<td>0.913</td>
<td>0.913</td>
<td>0.913</td>
</tr>
<tr>
<td>GDS2688</td>
<td>0.897</td>
<td>0.950</td>
<td>0.968</td>
<td>0.972</td>
<td>0.972</td>
<td>0.972</td>
<td>0.972</td>
<td>0.972</td>
<td>0.972</td>
</tr>
<tr>
<td>GDS5856</td>
<td>0.968</td>
<td>0.956</td>
<td>0.964</td>
<td>0.964</td>
<td>0.964</td>
<td>0.964</td>
<td>0.964</td>
<td>0.964</td>
<td>0.964</td>
</tr>
<tr>
<td>GDS5881</td>
<td>0.895</td>
<td>0.950</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
</tr>
</tbody>
</table>

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 ranks in which the results with the same CD are not significantly different (< 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) (Haielsellasse 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%), 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 time periods during 14 days of differentiation into embryoid bodies.

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 (GDS2666, (Hailesellasse Sene et al., 2009)) recently 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)

<table>
<thead>
<tr>
<th></th>
<th>GDS2666</th>
<th>GDS2667</th>
<th>GDS2668</th>
<th>GDS2669</th>
<th>GDS2671</th>
<th>GDS2672</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDS2666</td>
<td>/</td>
<td>0.915b</td>
<td>0.939b</td>
<td>0.901b</td>
<td>0.840b</td>
<td>0.818b</td>
</tr>
<tr>
<td>GDS2667</td>
<td>0.947a</td>
<td>/</td>
<td>0.949b</td>
<td>0.969b</td>
<td>0.869b</td>
<td>0.857b</td>
</tr>
<tr>
<td>GDS2668</td>
<td>0.980b</td>
<td>0.891b</td>
<td>/</td>
<td>0.893a</td>
<td>0.770b</td>
<td>0.804c</td>
</tr>
<tr>
<td>GDS2669</td>
<td>0.941c</td>
<td>0.954b</td>
<td>0.941a</td>
<td>/</td>
<td>0.828b</td>
<td>0.830b</td>
</tr>
<tr>
<td>GDS2671</td>
<td>0.955c</td>
<td>0.921b</td>
<td>0.953b</td>
<td>0.875c</td>
<td>/</td>
<td>0.711c</td>
</tr>
<tr>
<td>GDS2672</td>
<td>0.941c</td>
<td>0.960b</td>
<td>0.935c</td>
<td>0.909b</td>
<td>0.840a</td>
<td>/</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>G</th>
<th>N</th>
<th>Z</th>
<th>C</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-FC</td>
<td>0.883</td>
<td>1.000</td>
<td>0.905</td>
<td>0.983</td>
<td>0.943</td>
<td>2.875</td>
</tr>
<tr>
<td>PCA-AREA</td>
<td>0.883</td>
<td>1.000</td>
<td>0.917</td>
<td>0.971</td>
<td>0.929</td>
<td>2.875</td>
</tr>
<tr>
<td>PLS-AREA</td>
<td>0.950</td>
<td>0.933</td>
<td>0.905</td>
<td>0.950</td>
<td>0.935</td>
<td>3.125</td>
</tr>
<tr>
<td>MCE-euclid-AREA</td>
<td>0.983</td>
<td>0.700</td>
<td>0.905</td>
<td>0.583</td>
<td>0.793</td>
<td>4.000</td>
</tr>
<tr>
<td>MCE-euclid-FC</td>
<td>0.983</td>
<td>0.733</td>
<td>0.905</td>
<td>0.533</td>
<td>0.789</td>
<td>4.000</td>
</tr>
<tr>
<td>PCA-FC</td>
<td>0.883</td>
<td>0.867</td>
<td>0.857</td>
<td>0.983</td>
<td>0.898</td>
<td>4.125</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>G</th>
<th>N</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>/</td>
<td>1.000</td>
<td>0.976</td>
<td>1.000</td>
</tr>
<tr>
<td>G</td>
<td>0.933</td>
<td>/</td>
<td>1.000</td>
<td>0.950</td>
</tr>
<tr>
<td>N</td>
<td>0.850</td>
<td>0.883</td>
<td>/</td>
<td>0.817</td>
</tr>
<tr>
<td>Z</td>
<td>0.767</td>
<td>0.750</td>
<td>0.988</td>
<td>/</td>
</tr>
</tbody>
</table>

F, G, N, Z denote datasets (row labels) and test sets (column labels) that 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.
While Pathrecon is unsupervised, it only orders samples and does not provide an explicit model for staging of new samples. Because they know the samples come from different time points, expression profiles taken after 18 and 24 h might be very similar, but does not provide an explicit model for staging of new samples. 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 but does not provide an explicit model for staging of new samples. Poor predictions were obtained only for training and test data from different species.

While Pathrecon is unsupervised, it 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. This 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.

ACKNOWLEDGMENT

We thank Gad Shaulsky and Lucia Sacchi for useful comments and suggestions.

Funding: Fondazione Cariplo grant (2008–2006); European Commission FP7 grant (Health-F5-2010-242038); FIRB ITALBIONET grant (RBPR05ZK2Z); Slovenian Research Agency grants (P2-0209, J2-9699, L2-1112).

Conflict of Interest: none declared.

REFERENCES

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


