Title: Investigation of gene expressions in differentiated cell derived bone marrow stem cells during bone morphogenetic protein-4 treatments with Fourier transform infrared spectroscopy

Author names and affiliations: Jaber Zafari, Fatemeh Javani jouni, Ali Ahmadvand ,Parviz Abdolmaleki,Malihe Soodi, Rezvan Zendehe*

Jaber Zafari, Department of Toxicology and Pharmacology, Faculty of Pharmacy, Ahvaz Jundishapur University of medical sciences, Ahvaz, Iran (jaberzafari@yahoo.com).

Fatemeh Javani jouni, Department of Biophysics, Faculty of Biological Sciences.Tarbiat Modares University, Tehran, Iran (javani_f@yahoo.com).

Ali Ahmadvand, Computer Science and Informatics Department, EMORY University School of Medicine, Atlanta, GA, USA (ali.ahmadvand.66@gmail.com)

Parviz Abdolmaleki, Department of Biophysics, Faculty of Biological Sciences. Tarbiat Modares University, Tehran, Iran (parviz@modares.ac.ir).

Malihe Soodi, Department of Toxicology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran (soodi@modares.ac.ir).

Rezvan Zendehe, Department of Occupational Hygiene, School of Public Health, Shahid Beheshti University, Tehran, Iran.

Abstract:

A model was set up to predict the differentiation patterns based on the data extracted from FTIR spectroscopy. For this reason, bone marrow stem cells (BMSCs) were differentiated to primordial germ cells (PGCs). Changes in cellular macromolecules in the time of 0, 24, 48, 72, and 96 hours of differentiation, as different steps of the differentiation procedure were investigated by using FTIR spectroscopy. Also, the expression of pluripotency (Oct-4, Nanog and c-Myc) and specific genes (Mvh, Stella and Fragilis) were investigated by real-time PCR. However, the expression of genes in five steps of differentiation was predicted by FTIR spectroscopy. FTIR spectra showed changes in the template of band intensities at different differentiation steps. There are increasing changes in the stepwise differentiation procedure for the ratio area of CH₂, which is symmetric to CH₂ asymmetric stretching. An ensemble of expert methods, including regression tree (RT), boosting algorithm (BA), and generalized regression neural network (GRNN), was the best method to predict the gene expression by FTIR spectroscopy. In conclusion, the model was able to distinguish the pattern of different steps from cell differentiation by using some useful features extracted from FTIR spectra.

Keywords: FTIR, gene expression, regression models, differentiation

1. Introduction

Stem cells possess the exclusive ability to proliferate by self-renewal or to differentiate into tissues under the effect of suitable molecular prompts. There are various classes of stem cells, all varying in degree of pluripotency. Related to the developmental stage, stem cells are categorized into five main subcategories including those with totipotency, pluripotency, multipotency, unipotency, and oligopotency. Cells with totipotency can be transformed into all cell types, while those with pluripotency can be transformed into most cell types. Cells with multipotency can also be transformed into many cell types, while those with oligopotency can be transformed into a few cell types and those with unipotency into one cell type (1, 2). Mesenchymal stem cells (MSCs) are derived from several organs including the lungs, heart, adipose tissue, bone marrow, and liver (3-8). MSCs are capable of generating multiple mesenchymal cell lineages under specific culture conditions. Bone marrow mesenchymal stem cells (BMSCs) have potential multilineage and are used in several cell-based treatments.

Male germ cells are derived from a population of primordial germ cells (PGCs) set aside early during embryogenesis (9, 10). Cell proliferation, survival, and differentiation of PGCs are dependent on bone morphogenetic proteins (BMPs) (11). Bone morphogenetic protein 4 (BMP4) causes the expression of PGCs' specific genes such as Dpp3a (Stella), Fragilis, and mouse vasa homologue (Mvh). Western blots, quantitative polymerase chain reaction (qPCR), and immunocytochemistry (ICC) are nowadays used for differentiation assessments. Although such approaches are highly sensitive to specific cell characteristics, they have a destructive nature. This problem does not allow real-time assessments of cell functions(12).

Bio spectroscopy approaches were recently applied in studying of cells. This method has revealed characteristics of biological systems. Fourier transform infrared (FTIR) spectroscopy is one of the most powerful tools for detecting changes in cellular macromolecules: It provides information about the structures of carbohydrates, lipids, proteins, and nucleic acids, thereby allowing detection, identification, and quantification of changes in these macromolecules (13-16). This approach has been widely used to monitor biological processes involved in cancer investigation(17-19),apoptosis detection (20, 21), and stem cell research(22-24).

Biostatistics is a mathematical method for describing biological experimental data. This method is based on the probability theory and statistics that can make appropriate models to predict data. The number of biostatistics papers is increasing very fast, while the estimation of an experiment's results requires more sensitive methods(25, 26).

The aim of this study is to use a mathematical model to predict real-time PCR results based on the cellular FTIR spectroscopy of the differentiation of BMSCs into PGCs. This differentiation was investigated stepwise five times, while FTIR spectroscopy was employed to detect the changes in cellular macromolecules. Also, the expression of pluripotency (Oct-4, Nanog, and c-Myc) and specific genes (Mvh, Stella, and Fragilis) was investigated by using real-time PCR during differentiation. Four different static models were used to specify gene expressions by using an FTIR spectrum of cells. Models of different well-known regressions, including partial-least squares regression (PLSR), regression tree (RT), boosting algorithm (BA), and generalized regression neural network (GRNN), were initially used to predict the gene expression in selected samples based on the data extracted from FTIR profiles by using a sufficient sampling rate. We also used an ensemble of expert methods, including RT, BA, and GRNN, where the advantages of each model are combined together to get a better performance. The result of this project may provide new visions for the usage of FTIR in gene expression investigations.

2. Materials and Methods

2.1 Animal use

Male adult Vistar rats (aged=4-6 weeks old), derived from the original stocks obtained from Razi Laboratory (Tehran, Iran), were maintained under the standard conditions with free access to food and water at the animal facility of Tarbiat Modares University. To satisfy the moral concerns, all experiments were conducted in accordance with the guidelines set by the National Research Council (affiliated to Tarbiat Modares University).

2.2 Isolation, culture, and treatment of MSCs

The primary source for bone marrow stem cells (BMSCs) is bone marrow, although recent reports indicate BMSCs can be isolated from other sources such as peripheral blood (27), fat (28), skin (29), vasculature (30), and muscle (31). BMSCs are isolated, as described by Woodbury (32). In our study, eight to 10 week-old rats were killed, and their tibias and femurs were dissected and placed in phosphate buffered saline (PBS) on ice. Marrow was flushed out under sterile conditions by cutting two ends and using a syringe filled by 10 ml of α MEM (Invitrogen LT, Merelbeke, Belgium). Cells were seeded in α MEM supplemented with 20% of fetal bovine serum (FBS, Invitrogen LT, Merelbeke, Belgium), 100 U/ml of penicillin (Gibco, Germany), and 100 U/ml of streptomycin (Gibco, Germany) in a 75 cm² flask and incubated at 37°C with 5% humidity. After 24h, non-adherent cells were discarded and adherent cells were thoroughly washed twice with phosphate buffered saline (PBS). BMSCs were allowed to grow until 70%–80% confluence of the surface area of the flask. Confluent BMSCs were passaged and

plated out at 1:2 or 1:3 dilutions every five to six days by using 0.25% trypsin and 1 mM of EDTA (Invitrogen LT, Merelbeke, Belgium), and the media was changed every other day. MSCs were purified from other cells based on their adherent properties. Spindle-shaped cells were used from 5th passages for treatment (Figure 1). These cells were frozen in α MEM with 93% FBS and 7% DMSO (MERCK, Darmstadt, Germany) in liquid nitrogen (33). After four or five passages, when the density of the cultured BMSCs was approximately 5×10^4 cells cm⁻², the cells were ready to be treated with BMP4(32). BMP4 (25 ngml⁻¹; Chemicon, USA) was added to the 4th passage BMSCs at different times (0, 24, 48, 72, and 96 h), according to the method described by Mazaheri et al. (34). To study the change of macromolecules in the differentiation procedure, each time was considered as the first step (0, 24, 48, 72, and 96 h).

2.3 Quantitative analysis of gene expression

According to the manufacturer's recommendations, the total amount of RNA was extracted from the BMSCs treated with BMP4 in each step (0, 24, 48, 72, and 96 h) by using RNeasy Mini Kit (Qiagen, USA). Concentrations of RNA were determined by UV spectrophotometry (Eppendorff, Germany). The cDNAs were synthesized from 500 ng of DNAase-treated RNA samples with a Quantitect Reverse Transcription Kit by using oligo (dT) primers. The specific primers used for PCR reactions are listed in Table 1, and they were synthesized by Pishgam (Tehran, Iran). PCRs were performed by using Master Mix and Eva Green in an Applied Biosystems Step One™ thermal cycler (Applied Biosystems, USA). The PCR program started with an initial melting cycle to activate the polymerase for 5 min at 95 °C, followed by 40 cycles of melting (30 sec at 95°C), annealing (30 sec at 58°C), and extension (30 sec at 72°C). The quality of PCR reactions was confirmed by melting curve analysis. Efficiency was determined by using a standard curve for each gene (logarithmic dilution series of cDNA from the tests). For each sample, the reference gene (Gapdh) and target gene were amplified in the same run. The reference genes were approximately equal. The target genes were normalized to a reference gene and expressed relative to a reference gene as calibrator.

2.3 FTIR analysis

Cells were suspended with the trypsin (0.5 gL⁻¹)/EDTA (0.2 gL⁻¹) buffer and then washed three times in the normal saline solution (NaCl, 0.9%) to completely remove trypsin and medium. **The cells were counted and then 35 μ l of cell suspension containing 10^4 cells were deposited onto a**

BaF₂ window. The spectra were prepared from a layer cell suspension containing multiple cells. The window was placed into a drying vacuum to vaporize water (17). FTIR measurements were done in the transmission mode, where 100 scans were collected at a resolution of 4 cm⁻¹ for every wave number between 400 and 4000 cm⁻¹ by using the Thermo-Nicolet IR100 FTIR spectrometer. Two independent differentiation procedures were evaluated in five steps by FTIR assay with two replicates per test. Each spectrum is average from the four individual spectra of the same step. Each spectrum was baseline corrected and then normalized. The software of Essential FTIR version 2.00.045 was used to calculate the first and second derivatives of spectra.

2.4 Data analysis

Four different models were used to evaluate a suitable relationship between FTIR features and RNA patterns of stem cells in each step. They were PLSR, RT, BA, GRNN, and an ensemble of experts.

The observation (X) dataset consisted of 20 samples with 403 features extracted from the FTIR spectrum of cells and the gene expression profile was selected as the prediction (Y) matrix. Thus, the observation data was a complex matrix with a vector dimension of 403*20. To reduce the dimensions of this dataset, the principal component analysis (PCA) algorithm was applied.

2.5 Prediction models

PLSR, RT, and GRNN are well-known single methods in data processing. Ensemble methods use a combination of different models in their predictions. These methods have higher diversity and generalization capabilities than single models. They also have greater ability to handle small datasets that frequently occurred in biological studies. These methods include two major categories: ensembles of weak learners and ensembles of experts. The relationship between the observations and the prediction matrix was established by using training and testing sets in the prediction models. To perform these procedures, the dataset was divided into training and testing sets. Thus, the models were initially trained on training sets and then finally tested on testing sets. All the models were established by using MATLAB software (Version 4.0).

2.5.1 Partial least-squares regression (PLSR)

The partial least-squares (PLS) method is used for multivariate calibration. This model is based on the decomposition of observation and prediction datasets to latent variables. PLS is a regression technique for designing hidden variables in order to display the highest covariance among original observation and prediction matrices (X and Y).

2.5.2 Regression tree

Regression and classification tree are two main categories of decision trees. A decision tree is composed of root, internal nodes, and leaf nodes; it is a nonparametric model that implements the divide-and-conquer strategy. Each internal node is applied to a test function, which splits the dataset into different branches; the excellence of a split is measured by the mean square error from the estimated value. In each step, if the error is not tolerable, data attained to the N-th node are split further as long as the addition of the errors in the branches is remains (35).

2.5.3 Ensemble of weak estimators

The ensemble of weak estimators has two main categories: boosting and bagging algorithms. The boosting algorithm uses a set of weak learners such as a simple tree. To explain further, simple trees are small trees that produce accuracy a little better than random choice in classification problems. In this study, a boosting algorithm was used with multiple weak learners in the form of an ensemble of 50 weak regression trees based on a tradeoff between accuracy and computational burden cooperating with the boosting strategy (36-38).

2.5.4 Ensemble of experts

Each expert is a complex and strong estimator with high accuracy in a dataset. The challenging task in the ‘ensemble of experts’ method is to combine different predictions for each complex model. Selection and fusion were proposed to combine different results. In this study, the selection method was used to select the best classifier. In the selection method, for prediction of each test sample, one of the estimators in the ensemble was active and the other estimators were dormant. The test sample was predicted with the active or best estimator. The best estimator was selected on the basis of the 1-NN (KNN) algorithm. The KNN algorithm is a nonparametric method that uses a dissimilarity measure to find the K-nearest neighbour to each sample. The Euclidean distance was used as a dissimilarity measure. In this study, the 1-NN algorithm was used because our dataset was small. Three different models—RT, BA, and GRNN—were learned on the dataset separately. Then, for the estimation of test samples, we first found the nearest sample from those among the training samples by the 1-NN algorithm and afterward the best estimator in the ensemble in this training sample was selected. Next, the selected estimator was used for the prediction of test samples.

2.6 Statistical analysis

Data was presented as mean $\pm$ standard deviation (SD) and analysed by using the one-way repeated measure analysis of variance (ANOVA), followed by Tukey’s post-hoc test. The statistical software

package SPSS version 19.0 was used to perform the calculations. If $p < 0.05$, then the result was judged to be significant. For the real-time PCR data, the logarithmic values had to be converted to real values by raising 2 to the power of the $\Delta\Delta C_t$ value before any statistical analysis was performed.

3. Results

3.1 Gene expression in the differentiation procedure

Real-time PCR was performed to analyse the expression of a subset of pluripotency markers as well as primordial germ cell-specific genes. When BMSCs were differentiated into PGCs, the expressions of Oct-4, Nanog, and c-Myc were significantly decreased ($p \leq 0.05$) in all of the experimental groups (Figure 2).

The results demonstrated that BMP4 had a positive effect on early germ cell induction and caused regulation in the expression of the specific genes Mvh, Fragilis, and Stella in the derived BMSCs (Figure 3).

3.2 FTIR spectrum of cells in the differentiation procedure

BMSCs differentiated into PGCs were studied at five steps of 0, 24, 48, 72, and 96 h. The FTIR spectrum of cells is reported in Figure 4. The peak area band of cellular macromolecules was estimated in the regions of $1200\text{--}2000\text{ cm}^{-1}$ and $2500\text{--}3600\text{ cm}^{-1}$.

At least three areas of variation were found by comparing the ratios of bands :There are two peaks at about 2850 and 2920 cm^{-1} , which can be related to the CH_2 symmetric and CH_2 asymmetric stretching vibration bands, respectively (39). The ratio between the peak areas of CH_2 symmetric and CH_2 asymmetric stretching was increased by differentiation progression (Figure 5b). In recent years, CH_2 symmetric/ CH_2 asymmetric band was attributed to changes in the cell membrane (39). Our study might reflect a clear change in the membrane of cells in the differentiation procedure.

The vibration band at 1540 cm^{-1} is assigned to the β -sheet secondary structure of amide II(40, 41), and the peak at 1650 cm^{-1} is related to the secondary structure of amide I(42). The vibration bands of amide I and amide II increased from Stage 1 to Stage 5. The increases in the peak of amide I and amide II bands were calculated as the difference between the peak area of bands in each station and the same peak of those cells from pre-stage. The comparison between the increases in peak of amide I and amide II was carried out by using a paired student t-test. It is

obvious that the amid II bands increased more than amid I peaks ($p < 0.006$). However, there are increasing changes in the stepwise differentiation procedure for the ratio area of amide II to amide I (Figure 5a).

As shown in Figure 5c, the results obtained from these experiments show that the increase in the nucleic acids/amide II ratio (I_{1234}/I_{1540}) correlates with the steps of differentiation, and thereby produces a significant linear correlation.

3.3 Prediction Model

The problem of datasets with a high number of features is the ‘curse of dimensionality’. Therefore, we used a dimension reduction method to decrease the number of features. To reduce the dimensionality of the dataset, the PCA was used and the first 14 principle components were chosen. Figure 6a shows the percentage of variance explained in the variables (X) as a function of the first 14 principle components of our dataset. The results showed that the first two principal components had about 95% of all the variance. Figure 6b shows the percentage of variance explained in the response variable (Y) as a function of the first 14 principal components of our dataset. The results showed that the first seven principal components had about 95% of all the variance. Selection of the number of components should be carefully performed through a cross-validation method. Thus, we used the ‘leave one out’ method, which is an effective method for small datasets to train and test established models. In this method, if a dataset has N samples, then one sample will be used for testing and N-1 samples for training in each step. Hence, all the samples will be applied for testing and training, and, accordingly, a robust result will be obtained.

Among all methods used in this study, PLSR used the variance explained in Y; therefore, the first seven PLS components were used in the cross-validation section. Yet, the other methods were carried out by using the first two components in the cross-validation section. Figure 7 shows the measured mean square error (MSE) for all methods in the test dataset with different numbers of principle components.

As shown in Figure 7, the PLSR method with five components achieved the best results compared to other numbers of components. However, the ‘ensemble of experts’ method including the BA, RT, and GRNN algorithms, had better accuracy in the second component compared to all applied methods. Actually, the combining of these three methods created the best model by using lower components. Figure 8a shows the PLS weights for the first five components, while Figure 8b shows the PCA loadings for the two principal components. Actually, the PLS weights in the figure show the linear combinations of the original variables that defined the PLS components. Consequently, they indicated the degree of dependency of each component in the PLSR on the original variables, and they also showed the direction

of this dependency. In the same manner, the PCA loadings defined the degree of dependency of each component in the other methods on the original variables.

The ‘ensemble of experts’ method with two principal components was considered the best model. Figure 9 shows the real data and predicted values concerning 18 different samples of the best model. The vertical axis shows the values, while the horizontal axis shows the selected estimator in the selection part for each of the 18 samples.

4. Discussion

To identify and characterize BMSC differentiation, a number of molecular biology techniques are available. Among them, PCR and microarrays technologies, RNA in situ hybridization, and immunocytochemistry allow monitoring of the expression of specific genes. Even if these approaches make it possible to obtain detailed information on single differentiation events, non-invasive and rapid methods to characterize differentiation would be of great relevance and impact in biomedicine (15). With this aim, we explored the potential of FT-IR spectroscopy to monitor the gene expression in the differentiation process of BMSCs into PGCs. Indeed, among spectroscopic techniques, FT-IR and Raman spectroscopies(43) have proved to be powerful tools to investigate molecularly complex biological systems (44-46). In particular, important results on stem cell characterization and differentiation were recently obtained by the FT-IR spectroscopy implemented with multivariate analysis (47-49). Here we reported an FT-IR spectroscopy study to monitor stepwise changes in the responses of nucleic acids, lipids, carbohydrates, and proteins expressed by rat BMSCs during differentiation with commercial techniques monitoring DNA molecules before and after differentiation. With this technique, we obtained an infrared absorption spectrum from a limited number of cells, taken directly from their culture, in a few minutes(15).

The justification for the selected genes was that three of them (Nanog, Oct-4, and c-Myc) indicate the stemness while the others (Fragilis, Mvh, and Stella) indicate the degree of differentiation of the treated BMSCs into PGCs. Our results show that the treatment of cells with 25ngml^{-1} BMP4 caused a significant decrease in the expression of pluripotency genes as well as a significant increase in primordial germ-cell-specific genes when compared with the control group (Figure 2). The Nanog and Oct-4 genes are essential for maintaining the pluripotency of embryonic stem (ES) cells in early embryos (50). BMP4 augments the expression of the Ifitm3 (Fragilis) gene in the proximal epiblast, marking the first step in the germ line commitment; before this, Ifitm3 is weakly expressed throughout the epiblast. Soon after gastrulation begins, presumptive germ cells lose their responsiveness to BMP4, and a subset of Ifitm3⁺ cells, which have migrated to the extra-embryonic mesoderm start expresses the Dpp3a (Stella)

gene. These represent the first population of fetal germ cells (primordial germ cells, or PGCs) that can be identified based on their alkaline phosphatase activity. During this process, the expression of a gene associated with pluripotency, Pou5f1 (also known as Oct4), becomes restricted to the germ line: Pou5f1 expression is seen in the inner cell mass (ICM), epiblast, and then in PGCs, while it becomes repressed in somatic cell lineages. These are the key events known to occur during germ line specification (51).

FTIR spectroscopy is a measurement of the absorption of IR radiation by a sample. The IR spectral data of high polymers is usually interpreted in terms of the vibrations of a structurally repeating unit (52, 53). The repeat polypeptide and protein units give rise to nine characteristic IR absorption bands, namely amide A, B, and I–VII. Of these bands, amides I and II are the two most prominent vibrational bands of the protein backbone (53–55). The most sensitive spectral region to the protein's secondary structural components is the amide I band ($1700\text{--}1600\text{ cm}^{-1}$), which is due almost entirely to the C=O stretch vibrations of the peptide linkages (approximately 80%). There is close correlation between frequencies of amide I band components and each secondary structural element of the proteins. In turn, this is determined by the secondary structure adopted by the polypeptide chain, reflecting the backbone conformation and hydrogen-bonding pattern. The observed amide I band contours the proteins or polypeptides and consists of overlapping component bands, representing α helices, β sheets, turns, and random structures (56). The amide II band, in contrast, is derived mainly from in-plane NH bending (40–60% of the potential energy) and from CN stretching vibration (18–40%), showing much less protein conformational sensitivity than its amide I counterpart (53). Other amide vibrational bands are very complex, they depend on the details of the force field, the nature of side chains, and hydrogen bonding, which makes them impractical options for protein conformational studies (56).

Previous reports have suggested that the concentration of macromolecules affects the intensity of IR absorption (57, 58). In the five differentiation steps in this study, the FTIR pattern showed changes in the band intensities (Figure 4). Table 2 shows the FTIR absorption bands of several functional groups of biochemical components. As shown in Figure 5, the absorption bands at 2920 cm^{-1} and 1650 cm^{-1} became more than 2850 cm^{-1} and 1540 cm^{-1} with increases in the five steps of differentiation (0, 24, 48, 72, and 96 h), respectively. It is known that some molecules' contents increase gradually as the differentiation of BMSCs is transformed into specific cells (59, 60). Here it seems that the amide II molecule increased more than the amide I content ($p < 0.006$), which could be used as a biomarker to estimate the differentiation steps. Figure 5b shows that I_{2921}/I_{2850} increased in the stepwise differentiation from 0 to 96 h. Our results show the ratios of C–H stretching bands increased significantly in all the five steps of differentiation. This can be attributed to the change in lipid structures in the cell membrane as well as protein and nucleic acid expressions (61).

An analysis of the observed changes in the nucleic acids compared with the protein content after differentiation was conducted. The ratio between bands at 1234 cm^{-1} (assigned to nucleic acids of the PO_2 asymmetric stretching band) and 1540 cm^{-1} (amide II) was estimated. These relationships were significantly linear in different steps of differentiation. Moreover, there were high correlations between the area band of nucleic acids PO_2 /amide II and Mvh, Oct4, Nanog, and c-Myc gene expressions (Figure 10). From this result, it could be proposed that the ratio between nucleic acids of the PO_2 asymmetric stretching and amide II bands of the cellular FTIR spectrum would be a useful assessment for differentiation monitoring instead of DNA content evaluation through real-time PCR, as the latter is costly and time consuming. Finally, the nucleic acids/amide II ratio is inversely associated with the apoptotic index (62) and increases in the nucleic acids/amide II ratio (I_{1234}/I_{1540}) correlates with cell growth and differentiation.

To evaluate a suitable relationship between FTIR features and the DNA pattern of stem cells, four different models were used. These models were BA, an ensemble of experts, PLSR, and RT. In general, ensemble methods, such as the ensemble of experts and BA, have higher generalizations compared to methods that use just one estimator for prediction. This is based on the increases in diversity as a result of using different learning strategies in decision making. Therefore, final results are more reliable than other single estimator methods and so our final results did not drop very much with changes in datasets or features. The prediction model comprised two main steps: a pre-processing step and a predicting step. Our dataset, just like other datasets in this area, had the curse of dimensionality phenomena, whereby the feature space was very complex with a high dimension in which features were intensely correlated with one another. Therefore, the pre-processing step was carried out to prepare the dataset for the estimation step. PCA was used to decrease the complexity of the feature space. Finally, the best principal components were used for the estimation step. Figure 6 shows the information that was presented in the first 14 principal components extracted from our dataset. The results show that with by using the first seven principal components, all the information from the features could approximately be covered. In the prediction step, a dataset with new feature vectors, which was extracted with the PCA algorithm was used. We used the leave-one-out cross validation (LOOCV) algorithm, which uses the predicted residual sum of squares (PRESS) function for the prediction step. This method is rather time-consuming but suitable for small datasets such as ours, which one sample from the dataset is used for testing and the rest are used for training. This method was assessed for each sample and an average value was reported. By using this algorithm, all samples were used for both training and prediction, and a robust and reliable estimation was obtained.

5. Conclusion

Specification is a serious obstacle in differentiation procedure. This study suggests that modeling of infrared spectroscopy may represent a suitable proceeding for following up differentiation. We introduced an ensemble method comprising several mathematical and computerized methods which was able to predict differentiation patterns. This model was worked by using some useful features extracted from FTIR spectrum, such as the amide II band, that highly correlated with findings from common molecular methods such as real-time PCR. The output of the model was comparable with the real data generated by real time PCR. This model presents a fast and inexpensive method, which could be used to distinguish the stepwise differentiation.

6. Acknowledgements

This work was financially supported by grant No. 95815665 provided by the Iranian National Science Foundation (INSF) and support from the national committee for development of sciences and technologies of stem cells.

References

1. Smith AG. Embryo-derived stem cells: of mice and men. Annual review of cell and developmental biology. 2001;17(1):435-62.
2. Ramathal C, Pera RR, Turek P. Embryonic Stem Cells and the Germ Cell Lineage.
3. Beltrami AP, Barlucchi L, Torella D, Baker M, Limana F, Chimenti S, et al. Adult cardiac stem cells are multipotent and support myocardial regeneration. Cell. 2003;114(6):763-76.
4. Griffiths MJ, Bonnet D, Janes SM. Stem cells of the alveolar epithelium. The Lancet. 2005;366(9481):249-60.
5. Fraser JK, Wulur I, Alfonso Z, Hedrick MH. Fat tissue: an underappreciated source of stem cells for biotechnology. Trends in biotechnology. 2006;24(4):150-4.
6. Hass R, Kasper C, Bohm S, Jacobs R. Different populations and sources of human mesenchymal stem cells (MSC): a comparison of adult and neonatal tissue-derived MSC. Cell Commun Signal. 2011;9(1):12.
7. Shi S, Gronthos S. Bone Marrow Mesenchymal Stem Cells. Stem Cells in Craniofacial Development and Regeneration. 2013:223-39.
8. Niemeyer P, Fechner K, Milz S, Richter W, Suedkamp NP, Mehlhorn AT, et al. Comparison of mesenchymal stem cells from bone marrow and adipose tissue for bone regeneration in a critical size defect of the sheep tibia and the influence of platelet-rich plasma. Biomaterials. 2010;31(13):3572-9.
9. Lawson K, Hage W. Clonal analysis of the origin of primordial germ cells in the mouse. Germline development. 1994;165:68-84.
10. McLaren A. Germ and somatic cell lineages in the developing gonad. Molecular and cellular endocrinology. 2000;163(1):3-9.
11. Lawson KA, Dunn NR, Roelen BA, Zeinstra LM, Davis AM, Wright CV, et al. Bmp4 is required for the generation of primordial germ cells in the mouse embryo. Genes & Development. 1999;13(4):424-36.

12. Javani Jouni F, Abdolmaleki P, Behmanesh M, Movahedin M. An *in vitro* study of the impact of 4mT static magnetic field to modify the differentiation rate of rat bone marrow stem cells into primordial germ cells. *Differentiation*. 2014;87(5):230-7.
13. Gasper R, Mijatovic T, Bénard A, Derenne A, Kiss R, Goormaghtigh E. FTIR spectral signature of the effect of cardiotonic steroids with antitumoral properties on a prostate cancer cell line. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. 2010;1802(11):1087-94.
14. Ami D, Mereghetti P, Natalello A, Doglia SM. Fourier transform infrared microspectroscopy as a tool for embryonic stem cell studies. *Methodological Advances in the Culture, Manipulation and Utilization of Embryonic Stem Cells for Basic and Practical Applications InTech*, Rijeka, Croatia. 2011.
15. Ami D, Neri T, Natalello A, Mereghetti P, Doglia SM, Zanoni M, et al. Embryonic stem cell differentiation studied by FT-IR spectroscopy. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*. 2008;1783(1):98-106.
16. Carmona P, Rodríguez-Casado A, Alvarez I, De Miguel E, Toledano A. FTIR microspectroscopic analysis of the effects of certain drugs on oxidative stress and brain protein structure. *Biopolymers*. 2008;89(6):548-54.
17. Derenne A, Van Hemelryck V, Lamoral-Theys D, Kiss R, Goormaghtigh E. FTIR spectroscopy: A new valuable tool to classify the effects of polyphenolic compounds on cancer cells. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. 2012.
18. Zendehdel R, Masoudi-Nejad A, H Shirazi F. Patterns Prediction of Chemotherapy Sensitivity in Cancer Cell lines Using FTIR Spectrum, Neural Network and Principal Components Analysis. *Iranian Journal of Pharmaceutical Research*. 2012;11(2):401-10.
19. Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. 2004;1688(2):176-86.
20. Gasparri F, Muzio M. Monitoring of apoptosis of HL60 cells by Fourier-transform infrared spectroscopy. *Biochemical Journal*. 2003;369(Pt 2):239.
21. Di Giambattista L, Pozzi D, Grimaldi P, Gaudenzi S, Morrone S, Castellano AC. New marker of tumor cell death revealed by ATR-FTIR spectroscopy. *Analytical and bioanalytical chemistry*. 2011;399(8):2771-8.
22. Tanthanuch W, Thumanu K, Lorthongpanich C, Parnpai R, Heraud P. Neural differentiation of mouse embryonic stem cells studied by FTIR spectroscopy. *Journal of molecular Structure*. 2010;967(1):189-95.
23. Heraud P, Ng ES, Caine S, Yu QC, Hirst C, Mayberry R, et al. Fourier transform infrared microspectroscopy identifies early lineage commitment in differentiating human embryonic stem cells. *Stem Cell Research*. 2010;4(2):140-7.
24. Ami D, Natalello A, Mereghetti P, Neri T, Zanoni M, Monti M, et al. FT-IR spectroscopy supported by PCA-LDA analysis for the study of embryonic stem cell differentiation. *Journal of Spectroscopy*. 2010;24(1-2):89-97.
25. Boda K. Practical application of biostatistical methods in medical and biological research.
26. Voit E, Qi Z, Miller G. Steps of modeling complex biological systems. *Pharmacopsychiatry*. 2008;41(1):S78.
27. Kuznetsov SA, Mankani MH, Gronthos S, Satomura K, Bianco P, Robey PG. Circulating skeletal stem cells. *The Journal of cell biology*. 2001;153(5):1133-40.
28. Zuk PA, Zhu M, Mizuno H, Huang J, Futrell JW, Katz AJ, et al. Multilineage cells from human adipose tissue: implications for cell-based therapies. *Tissue engineering*. 2001;7(2):211-28.
29. Mizuno S, Glowacki J. Chondroinduction of human dermal fibroblasts by demineralized bone in three-dimensional culture. *Experimental cell research*. 1996;227(1):89-97.
30. Brighton CT, Lorch DG, Kupcha R, Reilly TM, Jones AR, ROBERT A WOODBURY I. The pericyte as a possible osteoblast progenitor cell. *Clinical orthopaedics and related research*. 1992;275:287-99.

31. Lee JY, Musgrave D, Pelinkovic D, Fukushima K, Cummins J, Usas A, et al. Effect of bone morphogenetic protein-2-expressing muscle-derived cells on healing of critical-sized bone defects in mice. *The Journal of Bone & Joint Surgery*. 2001;83(7):1032-9.
32. Sarvestani AS, Abdolmaleki P, Mowla SJ, Ghanati F, Heshmati E, Tavasoli Z, et al. Static magnetic fields aggravate the effects of ionizing radiation on cell cycle progression in bone marrow stem cells. *Micron*. 2010;41(2):101-4.
33. Marandi M, Mowla SJ, Tavallaei M, Yaghoobi MM, Jafarnejad SM. Proprotein convertases 1 and 2 (PC1 and PC2) are expressed in neurally differentiated rat bone marrow stromal stem cells (BMSCs). *Neuroscience letters*. 2007;420(3):198-203.
34. Mazaheri Z, Movahedin M, Rahbarizadeh F, Amanpour S. Different doses of bone morphogenetic protein 4 promote the expression of early germ cell-specific gene in bone marrow mesenchymal stem cells. *In Vitro Cellular & Developmental Biology-Animal*. 2011;47(8):521-5.
35. Rokach L. Data mining with decision trees: theory and applications: World scientific; 2008.
36. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*. 2005;27(2):83-5.
37. John Lu Z. The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2010;173(3):693-4.
38. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002;38(4):367-78.
39. Ke Y, Li Y, Wang ZY. The changes of Fourier transform infrared spectrum in rat brain. *Journal of forensic sciences*. 2012;57(3):794-8.
40. Baker MJ, Gazi E, Brown MD, Shanks JH, Gardner P, Clarke NW. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *British journal of cancer*. 2008;99(11):1859-66.
41. Zendejdel R, Masoudi-Nejad A, Mohammadzadeh J, H Shirazi F. Cisplatin resistant patterns in ovarian cell line using ftir and principle component analysis. *Iranian Journal of Pharmaceutical Research*. 2012;11(1):235-40.
42. Litvinov RI, Faizullin DA, Zuev YF, Weisel JW. The α -helix to β -sheet transition in stretched and compressed hydrated fibrin clots. *Biophysical journal*. 2012;103(5):1020-7.
43. Naumann D. FT-infrared and FT-Raman spectroscopy in biomedical research. *Applied Spectroscopy Reviews*. 2001;36(2-3):239-98.
44. Kretlow A, Wang Q, Kneipp J, Lasch P, Beekes M, Miller L, et al. FTIR-microspectroscopy of prion-infected nervous tissue. *Biochimica et Biophysica Acta (BBA)-Biomembranes*. 2006;1758(7):948-59.
45. Naumann D, Helm D, Labischinski H. Microbiological characterizations by FT-IR spectroscopy. *Nature*. 1991;351(6321):81-2.
46. Wood BR, Tait B, Mcnaughton D. Fourier transform infrared spectroscopy as a method for monitoring the molecular dynamics of lymphocyte activation. *Applied Spectroscopy*. 2000;54(3):353-9.
47. Bentley AJ, Nakamura T, Hammiche A, Pollock HM, Martin FL, Kinoshita S, et al. Characterization of human corneal stem cells by synchrotron infrared micro-spectroscopy. *Molecular vision*. 2007;13:237.
48. German MJ, Pollock HM, Zhao B, Tobin MJ, Hammiche A, Bentley A, et al. Characterization of putative stem cell populations in the cornea using synchrotron infrared microspectroscopy. *Investigative ophthalmology & visual science*. 2006;47(6):2417-21.
49. Krafft C, Salzer R, Seitz S, Ern C, Schieker M. Differentiation of individual human mesenchymal stem cells probed by FTIR microscopic imaging. *Analyst*. 2007;132(7):647-53.
50. Hattori N, Imao Y, Nishino K, Hattori N, Ohgane J, Yagi S, et al. Epigenetic regulation of Nanog gene in embryonic stem and trophoblast stem cells. *Genes to Cells*. 2007;12(3):387-96.
51. Nagano MC. In vitro gamete derivation from pluripotent stem cells: progress and perspective. *Biology of reproduction*. 2007;76(4):546-51.
52. Elliott A, Ambrose E. Structure of synthetic polypeptides. *Nature*. 1950;165(4206):921-2.

53. Krimm S, Bandekar J. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Advances in protein chemistry*. 1986;38:181-364.
54. Surewicz WK, Mantsch HH. New insight into protein secondary structure from resolution-enhanced infrared spectra. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*. 1988;952:115-30.
55. Susi H, Byler DM. Resolution-enhanced fourier transform infrared spectroscopy of enzymes. *Methods in enzymology*. 1986;130:290-311.
56. Kong J, Yu S. Fourier transform infrared spectroscopic analysis of protein secondary structures. *Acta Biochimica et Biophysica Sinica*. 2007;39(8):549-59.
57. Cakmak G, Togan I, Severcan F. 17 β -Estradiol induced compositional, structural and functional changes in rainbow trout liver, revealed by FT-IR spectroscopy: a comparative study with nonylphenol. *Aquatic toxicology*. 2006;77(1):53-63.
58. Bogomolny E, Argov S, Mordechai S, Huleihel M. Monitoring of viral cancer progression using FTIR microscopy: a comparative study of intact cells and tissues. *Biochimica et Biophysica Acta (BBA)-General Subjects*. 2008;1780(9):1038-46.
59. Liu ZJ, Zhuge Y, Velazquez OC. Trafficking and differentiation of mesenchymal stem cells. *Journal of cellular biochemistry*. 2009;106(6):984-91.
60. Shaffer A, Shapiro-Shelef M, Iwakoshi NN, Lee A-H, Qian S-B, Zhao H, et al. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity*. 2004;21(1):81-93.
61. Esau C, Kang X, Peralta E, Hanson E, Marcusson EG, Ravichandran LV, et al. MicroRNA-143 regulates adipocyte differentiation. *Journal of Biological Chemistry*. 2004;279(50):52361-5.
62. Gardner L, Lee L, Dang C. The c-Myc oncogenic transcription factor. *Encyclopedia of Cancer*. 2002;2.
63. Meade AD, Lyng FM, Knief P, Byrne HJ. Growth substrate induced functional changes elucidated by FTIR and Raman spectroscopy in in-vitro cultured human keratinocytes. *Analytical and bioanalytical chemistry*. 2007;387(5):1717-28.

Table 1 Primer used for real- time PCR

Accession Number	Gene	Product Length	Primer Sequences	Melt Temperature (°C)
NM_001009178	Oct4	192	Forward: TTCAGCCAGACAACCATC Reverse: CCTCACACGGTTCTCAATG	73.7
NM_001100781	Nanog	161	Forward: TTCTGAACCTGAGCTATAAGCAG Reverse: GAGAATAGCTGCAATGGATG	74.6
NM_012603	C-Myc	115	Forward: GAGAACAGTTGAAACACAACTCG Reverse: TCTGTCAGAAGGAACCGTTCTCC	75.2
NM_001077647	Mvh	155	Forward: GCTCTGGTTTCTGGAAAG Reverse: ATCTTGGTCATATTCATCTCTC	70.5
NM_030833	Fragilis	76	Forward: AGTGTCTCATGGTCATTATC Reverse: TTTGTGAAAGCGACTGAAAG	76.9
XM_576413	Stella	172	Forward: CTGCTCCATTTCTTCATC Reverse: TTGTTCTTGTGCGATTG	79.2
NM_017008.3	Gapdh	212	Forward: TGTGACTTCAACAGCAACTCCCAT Reverse: CTCTCTGCTCTCAGTATCCTTGC	85.9

Table 2 Major FTIR spectra peak and assignments (15, 39, 63)

Wave Number (cm ⁻¹)	Band Assignments	Functional Group
1234	PO ₂ ⁻ asymmetric stretching	Nucleic acid
1393	CH ₃ asymmetric Bending, COO ⁻ symmetric stretching	Methyl groups in proteins, fatty acids, and amino acids
1456	CH ₂ bending, CH ₃ symmetric bending	Mainly lipids and methyl groups in proteins
1540	N–H bending, C–N stretching	Amide II band
1650	C = O stretching	Amide I band
2850	CH ₂ symmetric stretching	Lipid
2920	CH ₂ asymmetric stretching	Lipid
2957	CH ₃ asymmetric stretching	Lipid

Figure 1.

Figure 2.

Figure3.

Figure 4.

Figure 5.

Figure6.

Figure 7.

Figure 8.

Figure 9.

Figure 10.

Figure legends:

Figure 1. Primary cell culture of bone marrow mesenchymal stem cells: (A), the population of heterogeneity of cells after 4 h culture; (B), the population of homogeneity of cells after five-passage culture

Figure 2. Real Time-PCR analysis. mRNA levels were normalized with respect to Gapdh, chosen as an internal control. Histograms showing mean expression values ($\pm$ SD, $n=3$; $P < 0.05$) that extracted from the profile of pluripotency genes expression in treated cell with BMP4 (25 ng/ml). Different letters refer to significant differences according to Tukey test.

Figure 3. Real Time-PCR analysis. mRNA levels were normalized with respect to Gapdh, chosen as an internal control. Histograms showing mean expression values ($\pm$ SD, $n=3$; $P < 0.05$) that extracted from profile of specific genes expression in treated cell with BMP4 (25 ng/ml). Different letters refer to significant differences according to Tukey test.

Figure 4. Average FTIR spectra of BMSCs differentiated into PGCs between (A), 2500-3600 cm^{-1} and (B), 1200-2000 cm^{-1} .

Figure 5. The intensity ratio of the bands (A), amide II /amide I, (B), CH₂ symmetric/CH₂ asymmetric stretching, (C), nucleic acid/Amide II in steps of differentiation.

Figure 6. (A), Percent of variance explained in the variables as a function of the number of components. (B), Percent of variance explained in the response variable.

Figure 7. MSE measure for different number of principal components

Figure 8. (A), PLS weights for 5 different PLS components. (B), PCA loadings for 2 different components.

Figure 9. Real and predicted values for ensemble of expert model with 2 principal components.

Figure 10. Correlation between the intensity ratio of the bands nucleic acid/Amide II and (A), c-Myc (B), Nanog (C), Oct4 (D), Mvh genes expression.