Journal of Biomedical Informatics 68 (2017) 104–111

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A reductionist approach to extract robust molecular markers from microarray data series – Isolating markers to track osseointegration

Anwesha Barik, Satarupa Banerjee, Santanu Dhara, Nishant Chakravorty *

School of Medical Science and Technology, Indian Institute of Technology, Kharagpur, West Bengal 721302, India

ARTICLE INFO

Article history:
Received 14 October 2016
Revised 3 February 2017
Accepted 10 March 2017
Available online 11 March 2017

Keywords:
Osseointegration
Microarray data
Statistical classifiers
Gene ontology
Surface texture

ABSTRACT

Complexities in the full genome expression studies hinder the extraction of tracker genes to analyze the course of biological events. In this study, we demonstrate the applications of supervised machine learning methods to reduce the irrelevance in microarray data series and thereby extract robust molecular markers to track biological processes. The methodology has been illustrated by analyzing whole genome expression studies on bone-implant integration (ossointegration). Being a biological process, osseointegration is known to leave a trail of genetic footprint during the course. In spite of existence of enormous amount of raw data in public repositories, researchers still do not have access to a panel of genes that can definitively track osseointegration. The results from our study revealed panels comprising of matrix metalloproteinases and collagen genes were able to track osseointegration on implant surfaces (MMP9 and COL1A2 on micro-textured; MMP12 and COL6A3 on superimposed nano-textured surfaces) with 100% classification accuracy, specificity and sensitivity. Further, our analysis showed the importance of the progression of the duration in establishment of the mechanical connection at bone-implant surface. The findings from this study are expected to be useful to researchers investigating osseointegration of novel implant materials especially at the early stage. The methodology demonstrated can be easily adapted by scientists in different fields to analyze large databases for other biological processes.

1. Introduction

Modern day molecular biology tools and techniques like microarray, gene expression profiling have provided researchers with a vast armory to decode the molecular mechanisms of different biological processes. Advances in medicine and related fields require an in-depth understanding of these mechanisms at cellular level to identify and characterize the disease condition. In some cases, confirmation of the progress of the improvement in the clinical condition during the course of therapy also has been made easy by these techniques. Analysis of the enormous amount of data generated through these techniques is the known to be the most time consuming task [1]. The inconvenience caused thereby raises the challenge to the analysts towards extracting the most desired information. Difficulties in finding gene annotations and relating them to literature references have made this a tedious job. As a result of this, the interpretation of the data analysis cannot be easily linked to the previous studies. Finally, the rarity of standardized protocol limits the scope of the data mining. There are very few, practically none such single tool which can perform the tasks all together, including database storage, data queries, statistical analysis, clustering, functional analysis, interrelation within the relevant cluster and interaction with public databases as well as experimental outcomes on the Internet [2]. In order to extract meaningful information from freely available datasets/data series we have designed a methodology that can be easily adapted by researchers in different fields to extract succinct information from existing datasets. The application of this methodology has been demonstrated on osseointegration – phenomenon of crucial clinical importance.

The functional success of the oral prosthetics is directly related to the extent of osseointegration of the implants. Conceptualized by Branemark, osseointegration can be described as the “direct structural and functional connection between ordered, living bone and the surface of a load-carrying implant” ensuring a long term clinical stability of the implants [3,4]. Researchers working in the field of implantology often need to validate the osseointegrative potential of materials developed when compared with the existing ones. Although histological studies on samples collected from the bone-implant interface are considered to be among the most reliable methods in experimental in vivo analysis, such techniques involve invasive techniques. Comparative expression profiling of whole genome either by microarray or selective genetic profiling

* Corresponding author.
E-mail address: nishant@smst.iitkgp.ernet.in (N. Chakravorty).

http://dx.doi.org/10.1016/j.jbi.2017.03.005
1532-0464/© 2017 Elsevier Inc. All rights reserved.

by Real Time-Polymerase Chain Reaction (qPCR) provides an alternative approach to understand the underlying interactive mechanisms of cells, factors and chemical signals [5,6].

In spite of the existence of such advanced techniques implantologists are still elusive regarding the complex genetic networks underlying osseointegration. The existence of >30,000 functional genes makes it very challenging to interpret the results from large scale gene expression data. This prompts the necessity to identify a panel of molecular markers comprising of a minimal number of genes that can be used by researchers to monitor osseointegration especially in the early phase. Complications in the analyzing the large microarray data are mainly arisen by the huge dimensionality of the comparative gene expression data and availability of relatively small number of samples [7,8]. To overcome this, we have adopted a reductionist approach in this study using various statistical classifiers in order to identify the key genes necessary. This approach directs the analyst towards a significant conclusion by sequentially removing the statistically irrelevant genes which, at the end, retains the most desirable set of features.

2. Materials and methods

2.1. Selection of test set and validation set

The freely accessible public repository of functional genomics data, Gene Expression Omnibus (GEO) was used to search for datasets in the field of osseointegration biology. The keyword(s) "implant stability or osseointegration" searched in the tool revealed 133 results. Since the purpose of this study was to investigate osseointegration in human, we excluded the non-human results and this reduced the number to 130 results. Furthermore, as we required series of datasets for our analyses rather than individual samples, we applied the filter "series". Upon applying this filter the results showed up only two studies conducted on humans (GSE41446 and GSE42288) and the eliminated the remaining 128 individual datasets. Thus the two datasets, GSE41446 and GSE42288 were included for further analysis. The dataset GSE41446 comprised of samples collected from 11 human participants following 3 and 7 days of implant fixation with two different types of oral implants namely TiOBlast (micro-roughened surface) and Osseospeed (micro-roughened with superimposed nano-scale topography-Sup-Nano) [9]. The study was specifically designed to investigate the influence of surface topography on the transcriptional regulation at the bone-implant interface. The second dataset (GSE42288) was available from a study performed on retrieved titanium implants with surface topographies exhibiting Sup-Nano features (Osseospeed) and microrough surface (TiOBlast) which were placed in the alveolar bone of 11 systemically healthy subjects and 10 smoking subjects and subsequently harvested at 3 and 7 days after placement [10]. Given the fact that the primary purpose of our study was to identify a gene cluster suitable to monitor early stages of osseointegration, the involvement of smokers and non-smokers as separate categories in the dataset GSE42288 was not relevant and therefore were combined. Considering the diverse nature of the target population in mind, GSE41446 was selected as the "test set" whereas the dataset GSE42288 was used as the "validation set".

2.2. Categorization of test set and validation set

The test set GSE41446 was divided into four groups which were named as Class 1 - samples collected from participants treated with Sup-Nano implants at day 3, Class 2 - samples collected from participants treated with Sup-Nano implants at day 7, Class 3 - samples collected from participants treated with micro-roughened implants at day 3, Class 4 - samples collected from participants treated with micro-roughened implants at day 7 (Fig. 1).

As mentioned earlier, the validation dataset was obtained from a study conducted by Thalji et al. wherein they investigated the influence of smoking on osseointegration and therefore involved three factors, namely: day of sample collection, smoking habit and surface modification, thereby categorizing them into eight groups. We combined the data for "smokers" and "non-smokers" under different categories and hence were able to categorize the dataset into four different classes. The resultant classes of the validation dataset GSE42288 were as Class 1 - samples collected from smoker and non-smoker participants treated with Sup-Nano implants at day 3, Class 2 - samples collected from smoker and non-smoker participants treated with Sup-Nano implants at day 7, Class 3 - samples collected from smoker and non-smoker participants treated with micro-roughened implants at day 3, Class 4 - samples collected from smoker and non-smoker participants treated with micro-roughened implants at day 7.

2.3. Extraction of Differentially Expressed (DE) genes

The classified datasets were processed using the interactive web tool - NCBI-GEO2R. This tool allows users to compare between two or more groups in a GEO series in order to identify DE genes. This was followed by extraction of p values, log FC (Fold Change) values and gene symbols of the DE genes. Subsequently, the statistically significant DE genes were identified by setting the cut-off values of p value and log FC values as <0.05 and $\geq\pm2$ respectively.

2.4. Isolation of subset of genes and validation

Keeping either the surface topography or the day of study constant, subset(s) of genes were selected by sequentially reducing the features using three classifiers namely Naïve-Bayes (NB), k-Nearest Neighbors (kNN), Support Vector Machine (SVM) at 10-fold cross validation using Orange Canvas v2.7 [11,12]. The classification efficiency for each of the reductions as well as classifiers was constantly monitored in order to maximize the classification accuracy, sensitivity, specificity and area under the curve and to minimize the Brier score. Zero Brier score denotes the maximum accuracy in the predictions. The subsets of genes selected from the test set were validated with the dataset GSE42288 by evaluating classification accuracy, sensitivity, specificity, area under the curve and Brier score. The flow chart of the detailed study protocol is presented in Fig. 1. The details of the genes selected were retrieved from www.genecards.org [13].

2.5. Gene Ontology (GO)

All the DE genes between day 7 and day 3 (considering only p-value < 0.05) on both of the surfaces (micro-roughened and Sup-Nano) were extracted through GEO2Enrichr, a browser extension and server app available within GEO [14]. The functional analysis of those genes was performed by Enrichr [15,16]. Results of the analysis were represented in tabular form to evaluate gene ontology parameters like biological process, molecular function, and cellular component [17,18]. Top five functions or components were selected for the further discussion. Furthermore, the molecular functions or cellular components or biological processes involving less than five genes were excluded to match the significant predicted factors with the true one. The GO analyses were also performed with the DE genes extracted at different cut-off levels for logarithm of FC i.e. 1 and 2 while keeping the level of significance same.

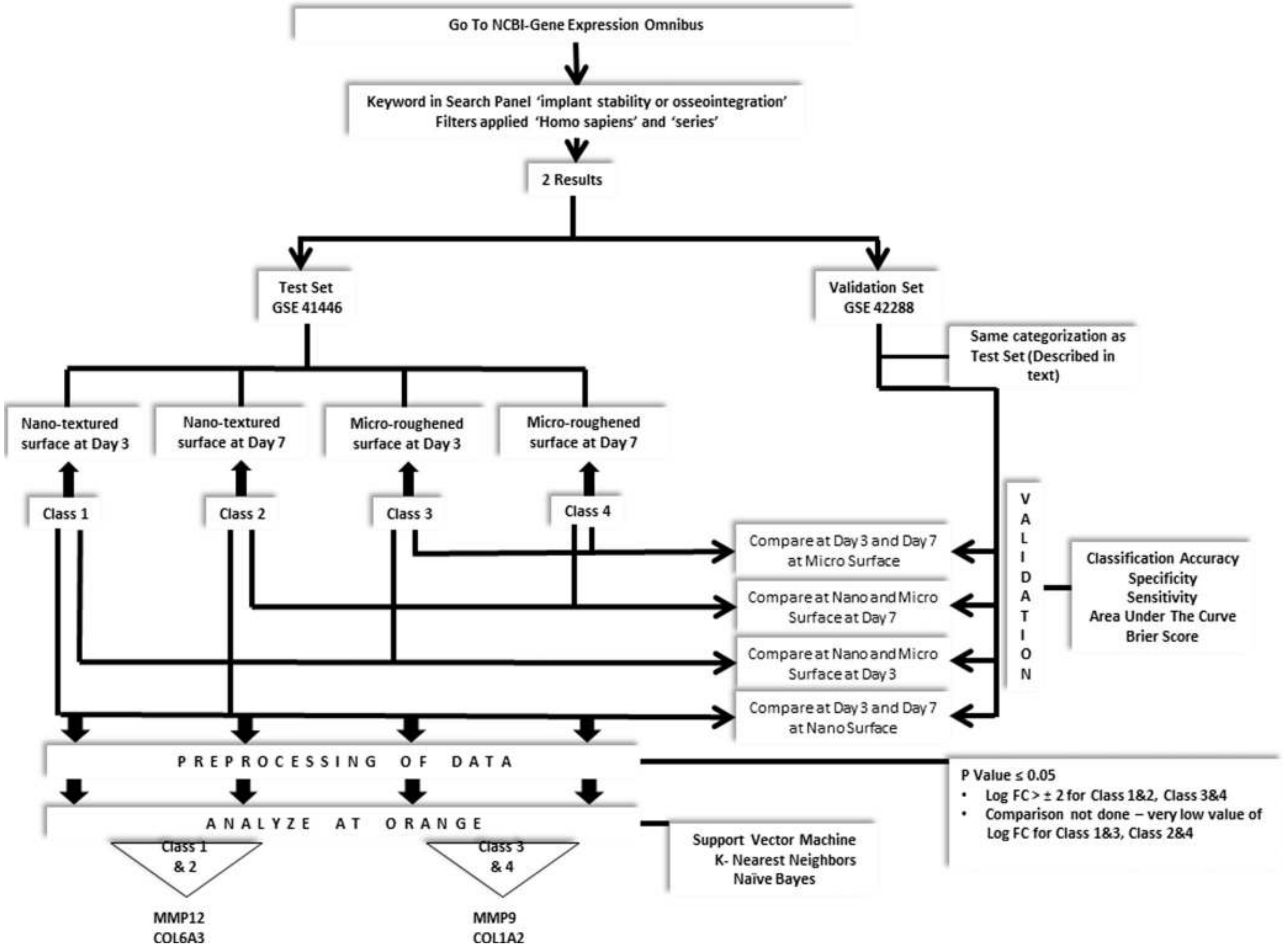
Go To NCBI-Gene Expression Omnibus
Keyword in Search Panel 'implant stability or osseointegration'
Filters applied 'Homo sapiens' and 'series'
2 Results
Test Set
GSE 41446
Validation Set
GSE 42288
Same categorization as Test Set (Described in text)
Nano-textured surface at Day 3
Nano-textured surface at Day 7
Micro-roughened surface at Day 3
Micro-roughened surface at Day 7
Class 1
Class 2
Class 3
Class 4
Compare at Day 3 and Day 7 at Micro Surface
Compare at Nano and Micro Surface at Day 7
Compare at Nano and Micro Surface at Day 3
Compare at Day 3 and Day 7 at Nano Surface
V A L I D A T I O N
Classification Accuracy
Specificity
Sensitivity
Area Under The Curve
Brier Score
PREPROCESSING OF DATA
P Value ≤ 0.05
Log FC > ± 2 for Class 1&2, Class 3&4
Comparison not done – very low value of Log FC for Class 1&3, Class 2&4
ANALYZE AT ORANGE
Support Vector Machine
K- Nearest Neighbors
Naïve Bayes
Class 1 & 2
Class 3 & 4
MMP12
COL6A3
MMP9
COL1A2

Fig. 1. Designed protocol for the selection of gene set from microarray data.

3. Results

3.1. Differentially expressed (DE) genes

After the classification of the test set GSE41446 was performed, the DE genes were extracted. While comparing class 1 and class 2 (Sup-Nano surfaces at day 3 vs. day 7), class 3 and class 4 (micro-roughened surfaces at day 3 vs. day 7), 13 (CHIT1, DCN, COL1A1, POSTN, COL3A1, COL6A3, COL1A2, LUM, CHI3L1, MMP7, MMP9, MMP12 and TM4SF19) and 22 (SLPI, GPNMB, POSTN, DCN, BGN, SPARC, ACP5, COL1A1, COL6A3, COL3A1, GM2A, COL1A2, LUM, CCL22, GAL, CHIT1, CTSK, CHI3L1, MMP7, MMP9, MMP12 and TM4SF19) genes listed out respectively on the basis of p value < 0.05 and log FC $\geq \pm 2$. All the DE genes identified following the GEO2R analysis showed upregulation in both of the comparisons except SLPI (Secretory Leucocyte Peptidase Inhibitor) which was found to be downregulated when compared between class 3 and class 4 (micro-roughened surfaces at day 3 vs. day 7). Similar analyses were performed to compare between class 1 and class 3 (Sup-Nano surfaces at day 3 vs. micro-roughened surfaces at day 3), class 2 and class 4 (Sup-Nano surfaces at day vs. micro-roughened surfaces at day 7). Although, none of the genes met the stringent criteria set for the assessments (p-value < 0.05 and log FC $\geq \pm 2$), however there was generally tendency towards upregulation on Sup-Nano surfaces compared to micro-roughened surfaces (172 genes upregulated and 148 genes downregulated at day 3; 182 genes upregulated and 124 genes downregulated and at day 7) when only p-value < 0.05 was considered as the cut-off criterion.

3.2. Isolation and validation of the most significant gene-set

Analysis using Orange Canvas v2.7 revealed that the genes MMP12 and COL6A3 taken together were able to track the progression of osseointegration on Sup-Nano surfaces between three and seven days with 100% classification accuracy, sensitivity and specificity for SVM and kNN classifiers. Although, Naïve-Bayes classifier showed slightly lower classification accuracy at (93.33%) and specificity at (81.82%), it was still high (>80%) and hence may be considered acceptable. Another subset of gene consisting of MMP9 and COL1A2 was found to track osseointegration on micro-roughened implant. In both of the cases, the lowest obtained Brier score denotes the highest accuracy in the probabilistic predictions done by kNN. The details may be found in Table 1.

3.3. Gene Ontology (GO)

The outcome of GO analysis without setting any cut-off for FC values (comparisons between day 7 and day 3 on both of the surfaces) performed by Enrichr is represented in Table 2 by selecting top five Biological Processes, Cellular Component, Molecular function all three and among them, those were excluded which involve less than five genes. The results from the other two GO analyses

Table 1
Selection and validation of subset of genes comparing the groups from where samples taken at 3 days and 7 days.

Surface texture	Groups compared	Selected combination of genes	Classifier used	Test set (GSE 41446)					Validation set (GSE 42288)				
				Classification accuracy	Sensitivity	Specificity	Area under the curve	Brier score	Classification accuracy	Sensitivity	Specificity	Area under the curve	Brier score
Sup-Nano (Osseospeed)	Day 3 Day 7	MMP12 COL6A3	Naive-Bayes	0.9333	1.0000	0.8182	1.0000	0.0581	1.0000	1.0000	1.0000	1.0000	0.0219
			k-nearest neighbor	1.0000	1.0000	1.0000	1.0000	0.0034	1.0000	1.0000	1.0000	1.0000	0.0000
			Support vector machine	1.0000	1.0000	1.0000	1.0000	0.0581	1.0000	1.0000	1.0000	1.0000	0.0282
Micro (TiOBlast)	Day 3 Day 7	MMP9 COL1A2	Naive-Bayes	1.0000	1.0000	1.0000	1.0000	0.0070	1.0000	1.0000	1.0000	1.0000	0.0133
			k-nearest neighbor	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0002
			Support vector machine	1.0000	1.0000	1.0000	1.0000	0.0325	1.0000	1.0000	1.0000	1.0000	0.0298

Table 2
GO analysis by comparing samples collected at Day 3 and Day 7 at different implant surfaces without FC cut-off.

Control	Experimental	Surface	Biological processes	Cellular component	Molecular function
Day 3	Day 7	Sup-Nano (Osseospeed)	Collagen metabolic process (GO:0032963)	Collagen trimer (GO:0005581)	Fibronectin binding (GO:0001968)
			Multicellular organismal macromolecule metabolic process (GO:0044259)	Cornified envelope (GO:0001533)	Collagen binding (GO:0005518)
			Collagen catabolic process (GO:0030574)	Lysosomal lumen (GO:0043202)	Extracellular matrix binding (GO:0050840)
			Multicellular organismal metabolic process (GO:0044236)	Vacuolar lumen (GO:0005775)	
			Multicellular organismal catabolic process (GO:0044243)		
		Micro (TiOBlast)	Collagen metabolic process (GO:0032963)	Fibrillar collagen trimer (GO:0005583)	Platelet-derived growth factor binding (GO:0048407)
			Multicellular organismal macromolecule metabolic process (GO:0044259)	Collagen trimer (GO:0005581)	Collagen binding (GO:0005518)
			Collagen catabolic process (GO:0030574)	Extracellular matrix (GO:0031012)	Extracellular matrix structural constituent (GO:0005201)
			Multicellular organismal metabolic process (GO:0044236)	Lysosomal lumen (GO:0043202)	Fibronectin binding (GO:0001968)
			Extracellular matrix disassembly (GO:0022617)	Cornified envelope (GO:0001533)	Proteoglycan binding (GO:0043394)

(cut-off values for log FC set at 1 and 2) as described in methods are presented in Table 3 and 4 respectively.

3.3.1. Biological processes

Identification of the biological functions associated with the early osseointegration showed collagen metabolic process (GO:0032963) to be the most enriched when day 7 samples were compared with day 3 ones, irrespective of the implant surface topography. The other processes enriched were mostly related to the structural organization of the extra-cellular matrix (ECM) (Table 2). When log FC cut-offs were selected at 1 and 2, the most enriched processes were found to be related to organization of ECM components and collagen metabolism obliterating the chemotactic activities while evaluating effect of days for both the surfaces (Tables 3 and 4).

3.3.2. Cellular component

Fibrillar collagen trimer (GO:0005583) with the highest enrichment score was identified as the most highlighted cellular component at day 7 (compared to day 3) on either implant surface. Other highly enriched components included collagen trimer (GO:0005581) and extracellular matrix (GO:0031012) for the comparisons made on micro-roughened surface (day 7 vs. day 3) (Table 2). Cut-off for log FC values at 1 resulted in enrichment of the components related to cell-matrix interaction at both types of implant surface (Table 3). Increase in the stringency of the criteria showed the collagen trimer and ECM as the highly enriched components associated with the process (Table 4).

3.3.3. Molecular function

Enrichment analysis of the molecular functions demonstrated the high-scoring terms to include fibronectin binding, collagen binding and PDGF binding. These types of attachments were related to the cell-matrix and cell-cell adhesion to the modified surfaces at early stage (Table 2). Setting of log FC value at 1 as a pre-determined criterion for evaluating the effect of days, collagen binding seemed to be important for both the implants but with some evidences of inflammatory activities specifically at micro-roughened surface (Table 3). Selection of more widely differentiated genes for GO analysis showed the structural assembly of ECM to be relevant at molecular level of functioning. Collagen binding showed some prospect in this aspect while studying the micro-roughened surface (Table 4).

4. Discussion

Modern day molecular technologies like DNA microarrays and qPCR profiling have vastly increased our understanding about the patterns of gene expression during biological responses in living systems. With the increasing use of such tools in implantology research, it has become possible for researchers to appreciate the gene expression profiles in different cells interacting with implant surfaces during osseointegration. However, the real messages hidden in these complex molecular interactions get dissipated only when the results are interpreted using appropriate statistical tools. In this study we have utilized statistical classifiers like SVM, NB and kNN to identify cluster of genes from whole microarray datasets available on the internet that may be able to monitor the course of osseointegration on implant surfaces. The scope of these three classifiers, kind of supervised learning method of classification fits well with the microarray like large datasets by featuring the ability to deal with large number of features and to identify the present outliers [19,20]. Probability of occurring error in the outcome prediction is very high when these classifiers are applied to extract smaller features from a very complex dataset. According

Table 3
GO analysis by comparing samples collected at Day 3 and Day 7 at different implant surfaces with log FC cut-off at 1.

Control	Experimental	Surface	Biological processes	Cellular component	Molecular function
Day 3	Day 7	Sup-Nano (Osseospeed)	Extracellular matrix organization (GO:0030198)	Extracellular vesicular exosome (GO:0070062)	Collagen binding (GO:0005518)
			Extracellular structure organization (GO:0043062)	Extracellular space (GO:0005615)	Extracellular matrix binding (GO:0050840)
			Extracellular matrix disassembly (GO:0022617)	Extracellular matrix (GO:0031012)	Extracellular matrix structural constituent (GO:0005201)
			Cellular component disassembly (GO:0022411)	Collagen trimer (GO:0005581)	
			Collagen metabolic process (GO:0032963)	Proteinaceous extracellular matrix (GO:0005578)	
		Micro (TiOBlast)	Extracellular matrix organization (GO:0030198)	Extracellular vesicular exosome (GO:0070062)	MHC class II protein complex binding (GO:0023026)
			Extracellular structure organization (GO:0043062)	Extracellular space (GO:0005615)	MHC protein complex binding (GO:0023023)
			Extracellular matrix disassembly (GO:0022617)	Extracellular matrix (GO:0031012)	Collagen binding (GO:0005518)
			Collagen metabolic process (GO:0032963)	Vacuolar part (GO:0044437)	Fibronectin binding (GO:0001968)
			Multicellular organismal macromolecule metabolic process (GO:0044259)	Cell surface (GO:0009986)	Chemokine activity (GO:0008009)

Table 4
GO Analysis by comparing samples collected at Day 3 and Day 7 at different implant surfaces with log FC cut-off at 2.

Control	Experimental	Surface	Biological processes	Cellular component	Molecular function
Day 3	Day 7	Sup-Nano (Osseospeed)	Extracellular matrix organization (GO:0030198) Extracellular structure organization (GO:0043062) Extracellular matrix disassembly (GO:0022617) Collagen catabolic process (GO:0030574) Multicellular organismal catabolic process (GO:0044243)	Extracellular matrix (GO:0031012) Proteinaceous extracellular matrix (GO:0005578) Extracellular space (GO:0005615) Collagen trimer (GO:0005581)	Extracellular matrix structural constituent (GO:0005201)
		Micro (TiOBlast)	Extracellular matrix organization (GO:0030198) Extracellular structure organization (GO:0043062) Extracellular matrix disassembly (GO:0022617) Collagen catabolic process (GO:0030574) Multicellular organismal catabolic process (GO:0044243)	Extracellular matrix (GO:0031012) Proteinaceous extracellular matrix (GO:0005578) Extracellular space (GO:0005615) Extracellular region (GO:0005576) Collagen trimer (GO:0005581)	Extracellular matrix structural constituent (GO:0005201) Collagen binding (GO:0005518)

to the vivid literature study, accurate prediction rule is constructed using 10-fold cross validation with least rate of prediction error [21]. The outcomes and interpretations of microarray data analysis also can be easily modulated by setting different FC and statistical cut-offs at different levels [22]. In order to obtain robust biological interpretations from comparative analysis between datasets, we kept stringent statistical cut-offs. The cut-off values for log FC and p-value were set as $\geq\pm2$ and <0.05 respectively, with an intent to isolate only those genes which vary widely and significantly in expression between the two conditions being compared.

Comparisons between days 3 and 7 on either micro- or super-imposed nano-textured demonstrated upregulation of all significant genes at day 7 except for SLPI (Secretory Leukocyte Peptidase Inhibitor). SLPI has an anti-inflammatory property. Complete absence of the gene is known to help in inflammation [23]. The downregulation of SLPI possibly helps in remodulation of the ECM. The role of MMP9 and MMP12 in the tissue remodeling through breakdown of ECM components on activation by extracellular proteinases validated the selection of those two genes. The rest two genes, COL6A3 functions by binding ECM components and COL1A2 has a role in fibril formation in the connective tissues. All the four genes, thus, aid in primary stability within a week of post-implantation period. It is pretty common to identify significant features by developing complex computer based algorithms [24,25] but this will need expertise and knowledge in computer coding. Therefore, the authors aimed towards framing an easy approach which is also adaptable for all end users to find significant information. The use of GEO2R and further classifying two disease conditions using the expression data with minimal number of genes and classifiers like SVM, kNN, NB is already gaining popularity in diseases like cancer. Table 5 shows classification accuracy of the genes selected using reductionist approach implementing GEO2R with SVM and other statistical classifiers in recent literatures with corresponding model efficiency and justifies the utility of the proposed methodology in current scenario for gene selection towards optimized disease and biological process classification. The uniqueness of this study is to harnessing biomarkers to differentiate a biological process rather than a disease and that also with 100% classification accuracy is first of its kind.

The significance of the outcome was further validated by biological interpretations drawn from GO analysis. Taking all the DE genes into account during GO analysis without any predetermined criterion, collagen metabolism was identified amongst the most enriched biological processes during early osseointegration irrespective of the surface modifications. Collagen, being the most abundant fibrous protein in mammalian ECM, provides the mechanical stability to the structure of tissue on association with the elastin microfibrils [30]. Owing to this, collagen metabolic and catabolic processes are considered equally important for cell differentiation on implant surfaces. Imbalance between these two processes results in alteration of structure of ECM, overall shape of tissue and optimal physical properties like mechanical loading, tensile strength. Among other top ten processes, organization of ECM and extracellular structure were identified. Such processes which help in wound repair by rearrangement, assembly or disassembly of the constituent parts resulting in reformation of original structure [31] seem important on nano-textured surface. However, the collagen fibril formation seems to be associated with both the surfaces. The collagen genes (Collagen Type I, III, XII) associated with the selected biological processes are mostly fibril forming collagen present in bone and connective tissues. Among them Type XII collagen in association with the Type I is known to modify the interaction between collagen fibrils and the neighboring matrix. In accordance with the biological processes, the topmost cellular component affected at the bone-implant interface was found to be the collagen trimer and ECM. In addition to these, var-

Table 5
Comparative analysis between methods reported in the literature.

Disease/process	Conditions classified	Gene selection process	Classification method used	Classification accuracy (%)	Reference
Osseointegration of Sup-Nano(Osseo speed) implants	Day 3 and Day 7 of implantation	GEO2R	SVM NB kNN	100 93.33 100	This study
Oral Cancer	Oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC)	GEO2R	SVM	98.89	Banerjee et al. [26]
Skin disease	Atopic dermatitis and normal skin	GEO2R	SVM	98	Ghosh et al. [27]
Neurodegenerative disease of eye	Age related Macular Degeneration (AMD) and normal condition	GEO2R	NB Decision Tree	89.66 66.67	Hao et al. [28]
Colon Cancer	Colon adenocarcinoma and normal condition	Minimum redundancy maximum relevance-Artificial Bee Colony (mRMR-ABC)	SVM	96.77	Alshamlan et al. [29]
Leukemia	Acute myeloid leukemia (AML) and Acute lymphoblastic leukemia (ALL)			100	
Lung Cancer	Stage I and Stage III early stage lung tumor			100	
Small Round Blue Cell Tumor (SRBCT)	Neuroblastoma (NB), Rhabdomyosarcoma (RMS), Burkitt lymphoma (BL) and the Ewing family of tumors (EWS)			100	
Lymphoma	Different categories of B cell malignancy			100	

ious studies reported increased osteoconduction and osseointegration on immobilization of type I collagen and other factors like PDGF (Platelet Derived Growth Factor), fibronectin on implant fixtures which validates the highly enriched molecular functions found from the analysis [32–34].

Fixing the cut-off for log fold-change values at 1 unfolded the similar role of ECM components at the bone-implant interface during the early stages. ECM structures itself as a network largely composed of the fibrous proteins collagen, elastin and associated-microfibrils, fibronectin, etc. Besides performing structural role in mammalian tissues, it regulates cell behaviors like proliferation, adhesion, migration, differentiation as well as death. Continuous remodeling of ECM modulated by the genetic products of Matrix Metalloproteinase (MMP) and Collagen (COL) genes influence these cell behaviors by maintenance of stem cell storage which play a vital role in damaged tissue restoration such as bone remodeling and wound healing [35]. Associated cellular components supported the outcome of biological processes which gets strengthened by the increased cellular interaction to ECM as revealed on the micro-roughened implant surface. Introduction of micro-roughness on the surface seems to induce binding of several proteins and factors at the surface along with some inflammatory consequences due to presence of foreign body. These findings corroborated with the discussion of the original work [9].

As a result of GO analysis based on log FC cut-off value as 2, the biological processes, molecular functions and the cellular components showed the same functions irrespective of the surface modifications as discussed previously. Increase in the cut-off levels resulted in the enrichment of only highly DE genes related to these functions-MMPs (7, 9, 12) and COL (1, 3, 6).

Interestingly, MMPs and COLs were the classes which were identified as the molecular markers to track the early osseointegration as discussed above. Based on the findings from statistical classifiers and GO analysis, it can be concluded that expression profiling of these genes only may give a direction towards the successful tracking of the process rather than full genome studies. Similar approach may be used to analyze other clinical conditions also.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgement

The authors are thankful to Indian Institute of Technology, Kharagpur for providing the infrastructural support. The author would like to thank Dr. Debdoot Sheet, Assistant Professor, Department of Electrical Engineering, Indian Institute of Technology, Kharagpur for his valuable insights.

References

[1] K.R. Hess, W. Zhang, K.A. Baggerly, D.N. Stivers, K.R. Coombes, Microarrays: handling the deluge of data and extracting reliable information, Trends Biotechnol. 19 (2001) 463–468.
[2] A. Fadiel, F. Naftolin, Microarray applications and challenges: a vast array of possibilities, Int. Arch. Biosci. 1 (2003) 111–1121.
[3] P.I. Branemark, Osseointegration and its experimental background, J. Prosthet. Dent. 50 (1983) 399–410.
[4] C.M. Stanford, J.C. Keller, The concept of osseointegration and bone matrix expression, Crit. Rev. Oral. Biol. Med. 2 (1991) 83–101.
[5] N. Chakravorty, S. Ivanovski, I. Prasadam, R. Crawford, A. Oloyede, Y. Xiao, The microRNA expression signature on modified titanium implant surfaces influences genetic mechanisms leading to osteogenic differentiation, Acta Biomater. 8 (2012) 3516–3523.
[6] N. Chakravorty, S. Hamlet, A. Jaiprakash, R. Crawford, A. Oloyede, M. Alfarsi, et al., Pro-osteogenic topographical cues promote early activation of

osteoprogenitor differentiation via enhanced TGFbeta, Wnt, and Notch signaling, Clin. Oral Implants Res. 25 (2014) 475–486.

[7] L.M. Fu, C.S. Fu-Liu, Evaluation of gene importance in microarray data based upon probability of selection, BMC Bioinform. 6 (2005) 1–11.

[8] L.M. Fu, E.S. Youn, Improving reliability of gene selection from microarray functional-genomics data, IEEE Trans. Inf. Technol. Biomed. 7 (3) (2003) 191–196.

[9] G.N. Thalji, S. Nares, L.F. Cooper, Early molecular assessment of osseointegration in humans, Clin. Oral Implants Res. 25 (2014) 1273–1285.

[10] G. Thalji, L.F. Cooper, S. Nares, Gene expression profiles of early implant adherent cells in smokers and nonsmokers, J. Oral Implantol. 41 (2015) 640–645.

[11] Y. Olvera-Carrillo, M. Van Bel, T. Van Hautegem, M. Fendrych, M. Huysmans, M. Simaskova, et al., A conserved core of programmed cell death indicator genes discriminates developmentally and environmentally induced programmed cell death in plants, Plant Physiol. 169 (2015) 2684–2699.

[12] M. Štajdohar, J. Demšar, Interactive network exploration with orange, J. Stat. Softw. 53 (2013) 1–24.

[13] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, D. Lancet, GeneCards: integrating information about genes, proteins and diseases, Trends Genet. 13 (1997) 163.

[14] G.W. Gundersen, M.R. Jones, A.D. Rouillard, Y. Kou, C.D. Monteiro, A.S. Feldmann, et al., GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions, Bioinformatics 31 (2015) 3060–3062.

[15] M.V. Kuleshov, M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucl. Acids Res. 44 (W1) (2016) W90–W97.

[16] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinform. 14 (2013) 128.

[17] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene ontology: tool for the unification of biology, Gene Ontol. Consort. Nat Genet. 25 (2000) 25–29.

[18] M.F. Ochs, A.J. Peterson, A. Kossenkov, G. Bidaut, Incorporation of gene ontology annotations to enhance microarray data analysis, Meth. Mol. Biol. 377 (2007) 243–254.

[19] C.D.A. Vanitha, D. Devaraj, M. Venkatesulu, Gene expression data classification using support vector machine and mutual information-based gene selection, Proc. Comput. Sci. 47 (2015) 13–21.

[20] K. Chitode, M. Nagori, A comparative study of microarray data analysis for cancer classification, Int. J. Comput. Appl. 81 (2013) 14–18.

[21] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. 99 (2002) 6562–6566.

[22] M.R. Dalman, A. Deeter, G. Nimishakavi, Z.-H. Duan, Fold change and p-value cutoffs significantly alter microarray interpretations, BMC Bioinform. 13 (2012) S11.

[23] N. Angelov, N. Moutsopoulos, M.-J. Jeong, S. Nares, G. Ashcroft, S.M. Wahl, Aberrant mucosal wound repair in the absence of secretory leukocyte protease inhibitor, Thromb. Haemost. 92 (2004) 288–297.

[24] W. Chu, Z. Ghahramani, F. Falciani, D.L. Wild, Biomarker discovery in microarray gene expression data with Gaussian processes, Bioinformatics 21 (2005) 3385–3393.

[25] H.-L. Huang, Y.-C. Wu, L.-J. Su, Y.-J. Huang, P. Charoenkwan, W.-L. Chen, et al., Discovery of prognostic biomarkers for predicting lung cancer metastasis using microarray and survival data, BMC Bioinform. 16 (2015) 54.

[26] S. Banerjee, A. Anura, J. Chakrabarty, S. Sengupta, J. Chatterjee, Identification and functional assessment of novel gene sets towards better understanding of dysplasia associated oral carcinogenesis, Gene Rep. 4 (2016) 131–138.

[27] D. Ghosh, L. Ding, U. Sivaprasad, E. Geh, J. Biagini Myers, J.A. Bernstein, et al., Multiple transcriptome data analysis reveals biologically relevant atopic dermatitis signature genes and pathways, PLoS One 10 (2016) e0144316.

[28] Y. Hao, G.M. Weiss, Gene selection from microarray data for age-related macular degeneration by data mining, in: Proceedings of the International Conference on Data Mining (DMIN): The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016, p. 125.

[29] H. Alshamlan, G. Badr, Y. Alohali, MRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling, BioMed Res. Int. 2015 (2015) 15.

[30] C. Frantz, K.M. Stewart, V.M. Weaver, The extracellular matrix at a glance, J. Cell Sci. 123 (2010) 4195–4200.

[31] A.J. Bailey, Changes in bone collagen with age and disease, J. Musculoskelet. Neuronal Interact. 2 (2002) 529–531.

[32] H.-Y. Ao, Y.-T. Xie, S.-B. Yang, X.-D. Wu, K. Li, X.-B. Zheng, et al., Covalently immobilised type I collagen facilitates osteoconduction and osseointegration of titanium coated implants, J. Orthop. Translat. 5 (2016) 16–25.

[33] E. Ortolani, M. Guerriero, A. Coli, A. Di Giannuario, G. Minniti, A. Polimeni, Effect of PDGF, IGF-1 and PRP on the implant osseointegration. An histological and immunohistochemical study in rabbits, Ann. Stomatol. (Roma) 5 (2014) 66–68.

[34] T.A. Petrie, C.D. Reyes, K.L. Burns, A.J. García, Simple application of fibronectin-mimetic coating enhances osseointegration of titanium implants, J. Cell Mol. Med. 13 (2009) 2602–2612.

[35] I. Stamenkovic, Extracellular matrix remodelling: the role of matrix metalloproteinases, J. Pathol. 200 (2003) 448–464.