Gene Reports 4 (2016) 131–138

Contents lists available at ScienceDirect

Gene Reports

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

Identification and functional assessment of novel gene sets towards better understanding of dysplasia associated oral carcinogenesis

Satarupa Banerjee [a,*], Anji Anura [a], Jitamanyu Chakrabarty [b], Sanghamitra Sengupta [c], Jyotirmoy Chatterjee [a]

[a] School of Medical Science and Technology, Indian Institute of Technology, Kharagpur 721302, India
[b] Department of Chemistry, National Institute of Technology Durgapur, India
[c] Department of Biochemistry, University of Calcutta, Kolkata, India

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 11 April 2016
Accepted 11 April 2016
Available online 23 April 2016

Keywords:
Oral epithelial dysplasia
Oral squamous cell carcinoma
Venn diagram
Gene sub-set selection

ABSTRACT

Oral epithelial dysplasia (OED) often precedes oral cancer. Understanding the underlying complex biological aspects of dysplasia associated oral carcinogenesis using important gene sets is thus important. Computation assisted gene set identification through different feature ranking and visualization techniques was therefore attempted in this study. Result suggested that, weighted support vector machine (SVM) could be useful for feature ranking and SVM for attribute selection. Alteration in keratinization, cell–cell communication and peptidase activity was the major affected phenomena, while extracellular matrix dynamics was also found to be hampered. During best gene subset identification, set of six genes could classify normal (NOM) and oral squamous cell carcinoma (OSCC) conditions and two sets comprising four genes in each could classify NOM and dysplastic (DYS) conditions with 100% sensitivity and specificity. A gene set, comprising 32 genes showed best efficacy of 94.12% sensitivity, 99.40% specificity and 98.89% accuracy during classification of DYS and OSCC.

1. Introduction

Oral epithelial dysplasia (OED) is often a step that precedes development of squamous cell carcinoma. It can either convert to oral squamous cell carcinoma (OSCC) or revert back to normal condition, if treated early. Till date there are no specific biomarkers which may be precisely utilized to assess malignant potentiality of oral precancers including OED. Histopathological evaluation of biopsy specimens still serves as gold standard for critical detection of grades of dysplasia and for predicting its malignant potentiality. However, the procedure lacks specificity and suffers from inter and/or intra-observer variability because of the paucity of unequivocal features of dysplasia that may be regarded as cardinal markers for accurate prediction of progression risks in oral pre-malignant disorders A recent review suggested that combination of selected biomarkers may be effective to address such problem (Banerjee and Chatterjee, 2015).

OED is a histopathological condition, where cytological and architectural characteristics of oral mucosa are altered. The role of OED in oral carcinogenesis is quite controversial. Some literature suggests that likelihood of malignant transformation of OED is significant (Al-Dakkak, 2010), while other studies have shown that there is no correlation between malignant potentiality and grade of dysplasia (Dost et al., 2014). In such circumstances, understanding the molecular progression of OED to OSCC is important and can no longer be avoided (Pitiyage et al., 2009). Semi-quantitative analysis of immunohistochemically stained tissue sections has been attempted to grade OED in precancers, (Anura et al., 2014) however, the procedures are still immature and have not yet been utilized in routine clinical practices. Comparative and quantitative assessments of histological grading and immunohistochemical expression of few key molecules to study the association between OED and OSCC were reported in few studies (Anura et al., 2014; Tabor et al., 2003). Molecular dissection of oral carcinogenesis has also been attempted through the analysis of proteome and deregulation of molecular network (Molinolo et al., 2009), but understanding the progression of OED to OSCC remains in its infancy. In silico analysis of microarray gene expression data is recently gaining interest for selection of candidate gene which may be subjected to gene ontology (GO) and functional enrichment analysis for understanding underlying molecular, biological and cellular activities of given gene sets and prioritizing candidate diagnostic indicators (Hindumathi et al., 2014).

In this study, an in-depth bioinformatic and statistical analyses of the microarray transcriptome were attempted to throw light on the process. Differentially expressed (DE) genes were primarily selected to dissect progression of OSCC through OED. Weighted support vector machine (SVM) was employed to select precise gene subset towards optimal classification of oral lesions, OED and OSCC. Venn diagram was implemented in visualization of complex association of different gene sets, to unearth their possible functional association (Kestler et al., 2005). The major aim of this cost-minimized strategy exercise is

Abbreviations: OED, oral epithelial dysplasia (OED); SVM, support vector machine; NOM, normal; OSCC, oral squamous cell carcinoma; DYS, dysplastic.

* Corresponding author. Tel.: +91 9474005265.
E-mail address: satarupa@smst.iitkgp.ernet.in (S. Banerjee).

http://dx.doi.org/10.1016/j.genrep.2016.04.007
2452-0144/

to select a novel gene sub-set which can modulate specificity and sensitivity of the classification task.

The main challenge of microarray data analysis includes high number of variables against a small sample size, from which meaningful gene sets have to be chosen which should classify the disease with maximum efficiency at optimum computational burden and diagnostic cost (Liu et al., 2011). Supervised machine learning classifiers such as Naïve Bayes (NB) (Wu et al., 2012) and k nearest neighbor (KNN) (Zhang and Deng, 2007) are commonly used for cancer microarray data classification in addition to support vector machine (SVM). In this study efficiency of these three classifiers were evaluated. Feature ranking and feature selection are routinely used to reduce data dimensionality and improve learning and predictive efficiencies. A recent study showed feature ranking utilizing weights from linear SVM yields better result even with non-identically distributed training and testing data [13]. ReleifF is a feature selection algorithm, which acts through filtering and is popularly used in cancer microarray data analysis. It randomly draws instances and after computing the nearest neighbors, weighs the feature. It comparatively provides higher weightage to the attribute which have higher differentiating ability of the instance from neighbors of other class (Wang and Makedon, 2004). Efficacy of feature selection algorithms such as weighted SVM was also evaluated in this study during gene selection. Several data visualization techniques are used in cancer microarray data towards knowledge discovery and class labeled data analysis [15]. Among them, VizRank is a simple gene set ranking technique, which works through utilizing visual projections of class labeled data. Here, we employed Radviz (Radial Coordinate visualization)(Novakova and Stepankova, 2009; Mramor et al., 2005) based gene identification with minimal gene numbers (three), to reduce computation cost, as well as to identify a subset of molecular criteria showing maximum efficacy which may potentially be implemented in routine diagnostics.

2. Materials and method

GSE30784 dataset was downloaded from Gene Expression Omnibus and used in this study, which consisted of 167 OSCC, 17 OED and 45 NOM samples. DE genes for each two class conditions were obtained using GEO2R (Barrett et al., 2013). The cut-off for gene selection was p value < 0.05 and log FC value ± 2. During 3 class disease classification, cut-off value was p value < 0.05 and F score more than 100.

Initially, all DE genes, both upregulated and downregulated gene sets were identified separately and then gene ontology (GO) analysis and pathway analysis for each gene set was performed using EnrichR (Chen et al., 2013) where common pathways as well as important biological process, cellular component and, molecular function were identified. In gene ontology (GO) analysis, when minimum of 5 genes were found to be present in any condition, was considered significant. When too many processes or functions were obtained, a threshold of combined score was considered and mentioned accordingly in the "Result and discussion" section. Pathway analysis was done using KEGG 2015 pathway. Common pathway and gene ontology analyses were performed with cut-off of combined score 25. The concept of combined score in EnrichR is to integrate both p value and z score with the formula $c = \log(p) \cdot z$ where c is the combined score, represented by p, p-value computed using the Fisher exact test, and z the z-score computed by assessing the deviation from the expected rank. Since Enrichr provides all three options for sorting enriched terms, combined score of 25, and p value < 0.005 were only considered (Chen et al., 2013). Venn diagram was prepared using three different gene subsets, as well as six sub-sets of up- and down-regulated genes to identify common and exclusive genes in each process using InteractiVenn (Heberle et al., 2015). Utilization of this method aided understanding of the complexity of association of both upregulated and downregulated gene sets. GO analysis and pathway analysis for each gene set were also again performed using EnrichR (Chen et al., 2013).

During specific gene subset selection for optimal disease classification, efficiency of different supervised classifiers namely SVM, KNN and Naïve Bayes was assessed using best features obtained through weighted SVM feature ranking method. Efficiency of another feature ranking techniques namely ReliefF was compared with weighted SVM and plotted accordingly with the best classifier obtained in the previous step, SVM. For selection of best feature subset, manual sequential feature reduction was carried out and optimal classification efficiency was evaluated at 10 fold cross-validation. The gene set obtained for NOM and OSCC was cross-validated in GSE9844 data set, which comprised of 12 NOM and 26 tongue OSCC samples. These analyses were performed in Orange 2.7 (Demšar et al., 2004). Visualization based classification by Radviz with minimal gene numbers (three) was also performed. Plots have been provided in the supplementary figure. Biological functions of the genes obtained in this study have also been mined from Genecard (Safran et al., 2010) and presented in

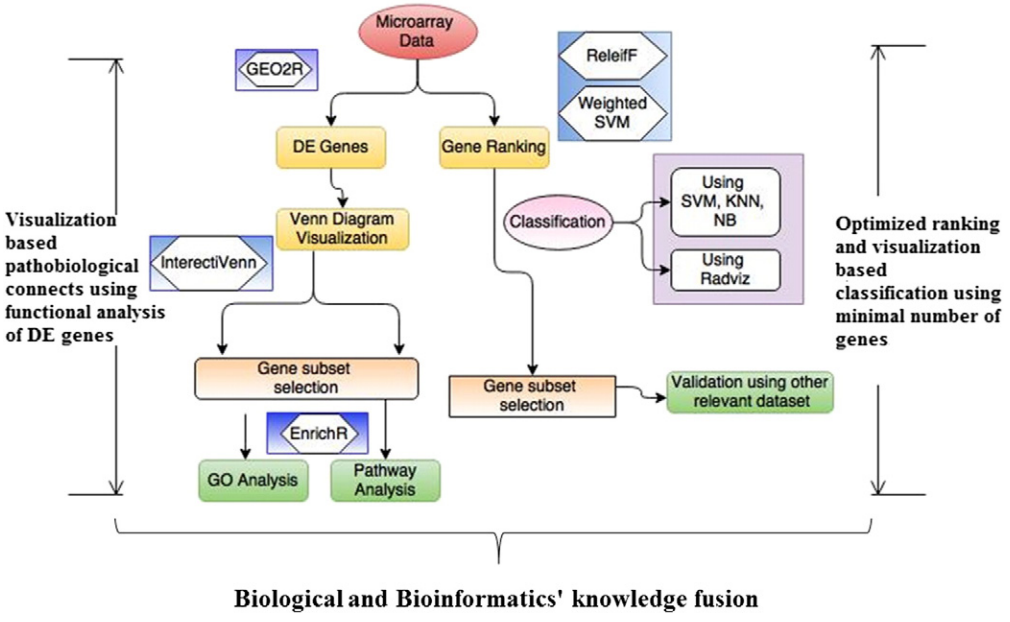

Fig. 1. Schematics of the methodology.