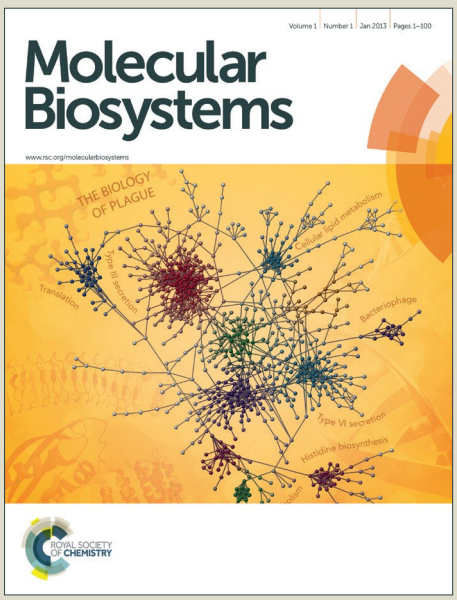
# Molecular BioSystems

Accepted Manuscript





ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

**Title:** A systematic approach to prioritize drug targets using machine learning, molecular descriptor-based classification model, and high-throughput screening of plant derived molecules: a case study in oral cancer

**Authors**

Vinay Randhawa[1,2] and Vishal Acharya[1,2,*]

**Author details**

[1]Functional Genomics and Complex Systems Laboratory, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Council of Scientific and Industrial Research, Palampur, Himachal Pradesh, India. [2]Academy of Scientific and Innovative Research (AcSIR), New Delhi, India.

E-mail addresses: Vinay Randhawa,vinay.plp@gmail.com; Vishal Acharya,vishal@ihbt.res.in, acharya.vishalacharya@gmail.com

*Correspondence: Phone: +91 1894-233339, Extn. 493; Fax: +91 1894-230433; Email addresses:vishal@ihbt.res.in, acharya.vishalacharya@gmail.com

**Abstract**

Systems-biology inspired identification of drug targets and machine learning-based screening of small molecules which modulate their activity have the potential to revolutionize modern drug discovery by complementing conventional methods.

To utilize the effectiveness of such pipeline, we first analyzed the dysregulated gene pairs between control and tumor samples and then implemented an ensemble-based feature selection approach to prioritize targets in oral squamous cell carcinoma (OSCC) for therapeutic exploration. Based on the structural information of known inhibitors of *CXCR4*—one of the best targets identified in this study—a feature selection was implemented for the identification of optimal structural features (molecular descriptor) based on which a classification model was generated. Furthermore, the *CXCR4*-centered descriptor-based classification model was finally utilized to screen a repository of plant derived small-molecules to obtain potential inhibitors.

Application of our methodology may assist effective selection of the best targets which may have previously been overlooked, that in turn will lead to the development of new oral cancer medications. The small molecules identified in this study can be ideal candidates for trials as potential novel anti-oral cancer agents. Importantly, distinct steps of this whole study may provide reference for the analysis of other complex human diseases.

**Keywords:** *CXCR4*, Feature selection, Logistic regression modeling, Machine learning, Oral squamous cell carcinoma, Plant derived molecules

## 1. Introduction

Early diagnosis of a disease can help to prevent its development and precise prognosis of a disease condition can avoid unnecessary treatments. A major focus in cancer research is therefore centered on the identification of disease-related genes (also called biomarkers). In the context of diseases, biomarkers are very crucial in serving as molecules for therapeutic intervention [1 2 3 4]; these biomarkers can also provide the basis for enhancing the prediction of patient-specific prognosis or therapeutic response [5 6]. While huge information is available about the identification of genes and developed methodologies can provide an opportunity for the selection of highly possible drug targets, it is still challenging to systematically integrate various resources to prioritize therapeutic drug targets. The classical approaches (such as linkage, candidate gene association, genome-wide association studies, etc.) are time consuming and even difficult to perform for identification of genes associated with complex diseases [7 8]. Furthermore, the genes identified by above mentioned approaches are usually not functionally related; therefore, these approaches offer limited usefulness in identifying specific genes those contribute to or are involved in complex diseases. Increasing evidences indicate that altered networks are the hallmarks of complex diseases, including cancers [9 10 11 12]. Instead of performing analysis for tens of thousands of gene comparisons, a network-based study limits the analysis to only few orders of magnitudes [13] ranging from hundreds [14 15] to even tens [16 17 18 19] of relevant genes. Considering these utilities, various network based approaches have been implemented to predict disease related genes [19 20]; however, most of these approaches still remain limited at the level of static regulation between genes rather than changes in the strength of gene-gene correlations across biological states. An approach of effective selection of the best targets from a large space will in turn lead to the production of only most successful drugs [21]. However, it is still a challenge to find the best druggable targets [22] that can affect the complex interaction networks [23]. Therefore, it will be highly helpful to develop and implement a systematic approach that can effectively integrate the available large-scale datasets and approaches for prioritizing the possible anti-cancer drug targets. Genes and proteins function cooperatively to regulate common biological processes by co-regulating each other [24]; therefore compared to traditional approaches, network-centered methodologies help in the better understanding of underlying complex interactions among genes. Because of these merits, these methods have been applied to prioritize

disease-associated genes in humans [25] and even for the identification of tissue-specific genes in plant systems [26].

Complex diseases are generally being caused by multiple aberrancies in the biological systems or networks rather than from changes in a single gene [27 28 29]. Given that diseases are often a consequence of perturbations in the strength of molecular interactions [25], approaches that focus on gene-gene correlations can be utilized to find drug targets, a method overlooked by various analyses that examine differential expressions only. In general, the best subset provides a higher accuracy in comparison to the original large dataset [30 31]; this kind of subsets have been proposed for molecular classification of various cancers [32 33 34]. Because there may be a large number of perturbed (dysregulated) genes across different biological states, the dimensionality problem still remain a major challenge. The ability of machine learning algorithms to reduce dimensionality and learn from the past examples to detect complex patterns from large data sets is particularly well-suited to medical applications; therefore, these techniques are extensively being applied for cancer prognosis and diagnosis [35 36 37 38 39]. Hence, machine learning approaches may be implemented to remove non-relevant genes in a supervised manner to find representative feature subsets that could satisfy a desired criterion.

Considering the central protein receptors identified by network-driven approaches, several successful attempts have been made to block the identified targets for therapeutic intervention [40 41]. In particular, there is a keen interest in inhibiting protein receptors by screening small-molecules from databases of natural compounds [42 43 44 45 46 47]. The molecules of plant origin possess various potential properties, including anti-diabetic, anti-tuberculosis [48], and even anti-cancer properties [49 50 51 52]. Additionally, plant based small-molecules also possess none or comparatively lesser side-effects; therefore, there is a growing interest in deriving these molecules for large-scale drug discovery. Similar to relevant gene-selection problem in genomics, the screening of selected molecules from large compound collections is also a major issue in cheminformatics research.

Because molecular structures can effectively be searched from drug-like libraries by the use of unique patterns (e.g., structural descriptors) of known structures [53], choosing appropriate structural patterns that could discriminate between molecules is the first and foremost priority. Furthermore, machine learning algorithms can solve the purpose by selecting only most relevant patterns in the structural dataset. In addition to their successful implementation in clinical

research for identification of genes, use of machine learning methods in cheminformatics research is growing in the last decades [54 55 56 57 58]. These methods had earlier been implemented for molecular structures classification problems [59 60 61 62 58 63 56 64] and also for screening of the molecules for therapeutic use [65 66]. These machine learning algorithms enable models to learn from data of known molecules (using molecular chemistry and structural information) and can successfully predict unknowns. Hence, using the recognized patterns of inhibitor datasets, the screening of similar compounds from large small-molecule databases may lead to large-scale identification of novel molecules for therapeutic intervention.

Oral cancer is one of the most common cancer worldwide with oral squamous cell carcinoma (OSCC) being the most common form which accounts for ~96% of oral cavity cancers [67]. Taking the analysis of high throughput data on OSCC as an example, herein a computational framework has been implemented which constitutes the following major steps: (1) prioritization of OSCC genes by identification of potentially dysregulated gene pairs and feature (gene) selection by logistic regression modeling, and (2) construction of a *CXCR4* (one of the most important drug target identified for OSCC) inhibitor-based classification model followed by screening of potent molecules from a repository of plant derived molecules (PDMs). To the best of our knowledge, in addition to prioritization of OSCC drug targets using ensemble methods, we also developed a classification model on the basis of known *CXCR4* inhibitors. The small molecules identified in the analysis can be ideal candidates for trials as potential novel anti-oral cancer agents.

## 2. Materials and Methods

### 2.1. Identification of candidate genes in oral squamous cell carcinoma (OSCC)

The R (http://www.r-project.org/) software package Variability Analysis in Networks (VAN) [68] was used for the identification of potentially dysregulated modules (comprising of a hub and all its interaction partners) in relation to OSCC disease phenotype. For this purpose, the gene expression dataset was analyzed in the context of protein-protein interactions (PPI) network. The gene expression profile of OSCC (comprising 355 tumor and 131 normal samples) was obtained from our previous study [69] while PPI data was compiled from various publicly available authoritative resources (Table 1). Detailed information of data preprocessing is provided in Supporting Information (Additional file 1: Supplementary Methods). During the analysis, a

threshold value of 30 was selected for defining a gene hub, and a total of 1000 permutations were performed to determine the Benjamimi-Hochberg (FDR) [70] adjusted p-value. The genes which had an FDR adjusted p-value less than 0.05 ($p<0.05$) were considered. The dysregulated healthy and cancerous networks were visualized by importing data into Cytoscape software package, version 3.0.1 [71].

**2.2. Disease enrichment analysis of the candidate genes**

The known cancer genes that are common to the candidate disease genes were used to evaluate the disease significance of obtained hub genes. The genes specific to OSCC were obtained from specialized databases which includes Head and Neck and Oral Cancer Database (HNOCDB) [72], The Oral Cancer Gene Database (OrCGDB) [73], and Oral Cancer Gene Database (Version II) [74] (all accessed in March, 2015). To broaden the scope of our study, a list of well curated and validated cancer genes (which are causally implicated in cancer) was obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [75] (accessed in Feb, 2015). The enrichment of candidate hub genes was estimated by comparing them to known cancer related genes using a hypergeometric test that is computed as follows:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where $N$, $K$ $n$ and, $k$ represent total number of gene expression profiles, number of known cancer associated genes, number of genes obtained in a sample, and number of candidate disease genes actually drawn in the experiment, respectively. $P$ is the statistical enrichment significance of the test.

Using an existing method [76], the random sampling was also performed to test the probability, where same number of known cancer genes was randomly picked in order to estimate whether these known cancer genes included in the previous results were statistically significant. Detailed information of method is provided in Supporting Information (Additional file 1: Supplementary Methods).

**2.3. Functional enrichment analysis of the candidate genes**

To obtain statistical enrichment of hub genes associated with specific biological processes and pathways annotated, gene ontology (GO) and pathway enrichment analysis were performed using ReactomePA [77] and GOstats [78] packages, respectively. To assess statistical enrichment, FDR-corrected hyper-geometric test p-values [79] were computed and overrepresented categories with enrichment p-values less than 0.001 ($p<0.001$) were considered as significant. The universe genes were defined as those which were present in the relevant background databases.

**2.4. Selection of target genes by ensemble based feature selection method**

For reducing the dimensionality in feature space, as a hybrid feature selection method, feature selection with random forest (RF) [80] was combined with elastic net logistic regression [81], where both approaches were implemented using R libraries "randomForest" [80] and "glmnet" [81]. To obtain a stable gene list, the feature selection procedure was performed for 1000 bootstraps (on ~60% of expression dataset). The genes that were present in more than a frequency threshold of 800 ($f>800$) iterations were finally selected. A similar procedure for feature selection was carried out for both RF and elastic net methods, and the overlapping genes (features) were considered candidate "oral cancer genes".

To confirm whether the identified oral cancer genes are disease biomarkers and could discriminate between healthy and tumor samples, four popular state-of-the-art supervised classification methods—conditional inference trees (party[82]), random forest and bagging ensemble using conditional inference trees (party[82]), bagging (ipred [83]), and support vector machine (SVM; e1071 [84])—were implemented. The classification models were constructed on the basis of 80% training dataset while evaluation was performed on the remaining 20% testing dataset. The classification power of each model was assessed using area under the receiver operating characteristic (ROC) and overall predictive accuracy in the ROCR software package [85]. In addition, leave-one-out cross-validation (LOOCV) strategy was also employed taking out one sample from the entire training data sets for test while keeping the remaining samples for training in each of $N$ rounds, where $N$ is the number of entire training data sets.

**2.5. Collection of *CXCR4* inhibitors and selection of best molecular descriptors**

CXC chemokine receptor 4 (*CXCR4*) is one of the important drug target identified in this analysis and is amenable to inhibition by small molecules; therefore, it was systematically probed for screening novel inhibitory small-molecules from a repository of PDMs.

All known *CXCR4* agonists, along with their inhibitory concentration (IC50), were collected from published data in the literature (Table 2), and their three-dimensional (3D) structures were drawn by means of Marvin Sketch 6.3 (ChemAxon, Budapest, Hungary) software. All molecules were subjected to geometry optimization using 500 steps of steepest descent method with MMFF94 force field [86]. The compound were assigned as inhibitors/active molecules if their IC50 (50% inhibition) values were less than 0.05 μM otherwise non-inhibitors (IC50≥0.05 μM); this resulted in a final data set which comprise of 81 inhibitors and 59 non-inhibitors.

A total of 1,444 one-dimensional (1D) and two-dimensional (2D) molecular descriptors were calculated for each molecule by the means of PaDEL-Descriptor software program [87]. As a preprocessing step, descriptors with more than 80% zero values and too small standard deviation values (<3%) were eliminated. Furthermore, Pearson correlation analysis (*r*) was employed (using "corrplot" library [88]) and the redundant/similar descriptors with correlation greater than 0.90 ($r$>0.90) were also removed. Similar to the selection of highly significant genes in relation to biological phenotype (section 2.4), frequency-based approach was used for obtaining the most relevant descriptors as the representative features of molecules. The statistical significance between average values of inhibitors and non-inhibitors descriptors was computed via Student's t-test. Multiple testing corrections were performed using the Benjamini & Hochberg method [70] for calculation of FDR adjusted p-value (q-value), and features with significant difference were retained (FDR adjusted p-values<0.05).

**2.6. Construction of a classification model and screening of plant derived molecules**

Random forest and bagging ensemble methods have been implemented utilizing conditional inference trees as base learners. In the section, a combination of two different classifiers, random forest and bagging, was used to create a classification model for best descriptors by the means of party [82] software package. The classification model was constructed using 80% training data set and tested for classification/prediction performance on rest of the 20% testing dataset. To validate the robustness, a 5-fold cross-validation scheme was employed, where training and testing were carried out five times in such a way that each time one set was used for testing and the remaining (n-1) sets for training. The best feature descriptors selected from each model were used for training the whole dataset to generate the final classification model. The sensitivity, specificity, overall predictive accuracy, and Matthew's correlation coefficient (MCC) were calculated for each test dataset in our 5-fold cross validation to test the performance of each

model. The classification model constructed using the whole dataset was also validated on an external independent data set which comprise of 39 active compounds against *CXCR4* obtained from DUD-E website (http://dude.docking.org/) [89].

The *CXCR4*-centered molecular descriptor-based classifier model developed was used for classifying molecules and screening of drug-relevant compounds from PDMs retrieved from two major Himalayan medicinal plant databases: SerpentinaDB [90] and Phytochemica [91]. The R ChemmineR [92] software package was used to cluster significant molecules into their discrete similarity groups. Furthermore, a maximum common substructure search was performed using the flexible common substructure algorithm [93] to identify potentially representative scaffolds in the clustered molecules.

**2.7. ADME/T prediction and molecular docking studies**

Absorption, distribution, metabolism, and excretion/toxicity (ADME/T) profiles of the top scored molecules were assessed from their respective databases (SerpentinaDB [90] and Phytochemica [91]), where computational prediction of pharmacokinetic properties were performed using ADMET descriptors in Discovery Studio v 4.0 (Accelrys, San Diego, USA). These molecules were also checked for their physicochemical properties by FAF-Drugs3 [94] web-server.

Furthermore, to obtain insights into structural interactions of the screened molecules, computational molecular docking studies were performed. The x-ray crystal structure of human chemokine *CXCR4* receptor in complex with isothiourea derivative (IT1t) inhibitor [95] (PDB ID: 3OE6; resolution: 3.20) was retrieved from the Protein Data Bank (PDB) [96]. The molecule (also called ligand) and receptor preparations for docking were performed by means of the Autodock Tools 1.5.4 software package [97] using our previously established protocol [43]. For software standardization, IT1t from co-crystallized complex was first extracted and then re-docked to its corresponding binding site (active site) using AutoDock Vina v 1.1.2 package [98]. For selecting the best PDMs, the bound IT1t and other non-protein molecules were removed from protein structure and molecules were docked into the active site. The active site was defined on the basis of bound IT1t inhibitor in crystal structure of *CXCR4*. Molecular interactions between protein and ligands were predicted using LigPlot$^{+}$ v 1.4.3 software [99] and molecular rendering was performed by means of the PyMOL software (PyMOL Molecular Graphics System, Version

1.5.0.1, Schrödinger, LLC). All computations were carried out on a 12-core HPZ600 workstation running Ubuntu 12.04 operating system.

## 3. Results and Discussion

The computational procedure implemented in this work is divided into two major steps: (1) prioritization of OSCC genes by identification of potentially dysregulated gene pairs and feature (gene) selection by logistic regression modeling, and (2) construction and validation of a *CXCR4* centered molecular descriptor-based classification model, followed by screening of potent inhibitory molecules from a compiled repository of PDMs.

### 3.1. Identification of network-level perturbations in healthy and tumor samples

It was assumed that those correlated gene pairs that potentially perturbed between health and cancer conditions possibly present the most significant genes associated to a disease. In this work, the genes that simultaneously possess low probability values (FDR adjusted p-value<0.05) and high log-fold change (2-fold higher or lower differential expression) were selected as signature genes on the basis of our earlier study [69] which identified 1,652 genes (1,052 over-expressed and 600 under-expressed) differentially expressed in OSCC tumor samples compared to their healthy counterparts (Additional file 2; Supplementary Table 1 and Supplementary Table 2). We selected DEGs because the combination of p-value and fold change criterion typically results in more biologically meaningful sets of genes [100 101 102]. Principal component analysis (PCA) statistical test is a technique for visualizing high dimensional data that reduces the dimensionality of multivariate data while retaining most of the variance, thus making data analysis and interpretation easy [103]. To visualize the overall expression patterns, we performed PCA on all DEGs used in this study to examine data in a two-dimensional plane. A clear separation between OSCC and normal groups was observed (Figure 1) with few outliers; this result indicates that normal and cancerous tissues had unique distinguishable expression profiles marked by different colors.

The hubs—proteins involved in many interactions—are thought to be candidate drivers and are also frequently observed among existing cancer therapeutic targets [104]. Therefore, identification of perturbed modules—hubs and all its interaction partners—are important for understanding the regulatory networks in a disease progression. Of all the total DEGs evaluated, 48 hub genes (in the respective modules) showed significant differences (p-value<0.05; FDR) in

the average gene expression correlation with respect to their interaction partners in healthy and cancerous state; therefore, they were potentially represented as dysregulated network markers in healthy and tumor tissues (Figure 2). This figure indicates a disruption in the coordination of gene expression among hubs and their interaction partners in normal (Figure 2A) and cancerous (Figure 2B) networks assigned by the changes in edge (connection) color. The significant changes in tumor samples indicate that genes with strongly altered connections can play a major role in cancer.

**3.2. The candidate hub genes are significantly enriched in oral cancer**

Among the 48 hub genes identified, 12 genes—*ANXA1* [105], *COL1A2* [106], *CXCR4* [107], *EGFR* [108], *FOS* [109], *ICAM1* [110], *IL6* [111], *MMP9* [112], *SERPINE1* [113], *STAT1* [114], *STAT3* [115], and *TGFB1* [116]—are well known predictive biomarkers or OSCC drug targets. There were a total 11 genes that were common between identified 48 hub genes and 547 COSMIC-obtained cancer genes. Of the whole background dataset of 1,652 genes analyzed, there were only 77 genes whose somatic mutations are implicated in cancer. To investigate whether these 11 genes could have obtained randomly, their enrichment was evaluated by hypergeometric test. A significant p-value ($p=3.5\times10^{-6}$) against $10^5$ random simulations (average p-value: $2.1\times10^{-1}$) was obtained; this result indicated that identified hub genes are enriched among known cancer related genes (rather than generated by chance). Similarly, there were 12 common genes identified between obtained hub genes in this study and 488 OSCC-specific genes (which were retrieved from HNOCDB, OrCGDB, and Oral Cancer Gene Database [Version II]). Of the whole background dataset of 1,652 genes analyzed, there were only 110 genes that were related to oral cancer. A significant p-value ($p=1.6\times10^{-5}$) was obtained against random simulations (average p-value: $1.7\times10^{-1}$); this observation is also an indicative of significant enrichment of hub genes among known OSCC genes.

Hub genes are thought to be candidate driver genes of a set of genes; therefore, elucidation of hub genes associated GO terms and pathways provide insight into the altered mechanisms in a diseased condition. The resulting GO term list returned by GOstats [78] was large and highly redundant, and was therefore difficult to analyze. REViGO web server (http://revigo.irb.hr/) [117] was implemented to summarize this long list. The terms with SimRel semantic similarity [118] of 0.5 were clustered to obtain a single representative term for each of the

identified clusters; this GO term summarization identified a total of 65 representative terms (Table 3.3).

ReVIGO GO summarization indicated that highly non-redundant representative terms with dispensability score ≤0.05 included representative terms associated to *immune system process* [119] (GO: GO:0002376, $p=1.01\times10^{-7}$), *nitric oxide metabolic process* [120] (GO:0046209, $p= 5.64\times10^{-5}$), *positive regulation of cellular component movement* [121] (GO:0051272, $p= 9.50\times10^{-19}$), *response to transforming growth factor beta* [122] (GO:0071559, $p= 1.34\times10^{-12}$), and *leukocyte cell-cell adhesion* [123] (GO:0007159, $p= 1.72\times10^{-7}$). The identified GO categories were significantly over-represented relative to their frequency in a randomized sample of expressed transcripts. Literature analysis revealed that the majority of these significantly enriched BP terms were related to molecular mechanisms associated with cancer associated processes. Other significant representative GO BP terms are summarized in supplementary table (Additional file 2, Supplementary Table 3). Most of these identified GO terms seem to be particularly interesting with respect to their well known role in cancer.

To further clarify the functional mechanism at molecular level, pathway enrichment analysis was performed on the basis of Reactome pathway database [124]. Pathway over-representation analysis presented *syndecan interactions* (ID:3000170; $p=2.23\times10^{-10}$) [125], *extracellular matrix organization* (ID:1474244; $p=2.23\times10^{-10}$) [126 127], *integrin cell surface interactions* (ID:216083; $p=2.19\times10^{-9}$) [128], *hemostasis* (ID:109582; $p=2.28\times10^{-9}$) [129], and *ECM proteoglycans* (ID:3000178; $p=9.63\times10^{-9}$) [127] as the top most important pathways over-represented in hub genes. These over-represented pathways are known to mediate various cancer associated processes which include cell growth and development [125], cell adhesion [128], drug resistance [126 127], and homeostasis [129]. Overall, these analyses altogether suggested that the hub genes mediate perturbations in various biological processes and pathways in the cancerous state.

**3.3. Identification of oral cancer targets using ensemble-based feature selection**

Some of the hub genes selected may be irrelevant to the trait of interest; therefore, only the subset of informative genes was probed further. Generally, hybrid feature selection approaches are effective in identification of the key genes that are associated with disease diagnosis or prognosis [39 130 131]; therefore, a practice of considering overlapped feature lists provides a promising approach [39]. As an ensemble approach, we considered the overlapped genes that were obtained in both feature selection methods (elastic net and RF) to obtain a more reliable subset of

genes. This combined criterion of feature selection reduced the feature vector space from 48 hubs to five genes—*ICAM1* (intercellular adhesion molecule 1), *ITGB1* (integrin, β 1), *CXCR4* (CXC chemokine receptor 4), *PTK2* (protein tyrosine kinase 2), and *COL1A2* (collagen, type I, α 2)—which were presumed to be highly related to OSCC. A detailed and systematic literature search inferred that the selected genes play major roles in oral cancer development and progression (*ICAM1*[132 133 134 135 136 137 138], *ITGB1* [139], *CXCR4* [140 141 142 143 144 145 146 147 148 149 150 151 152 153], *PTK2* [154 155], and *COL1A2* [156 157 158 159 160 161]). A study by Usami et al. [162] found *ICAM1* to play an important role in oral cancer progression angiogenesis, tumor invasion, and lymph node metastasis, cell adhesion. In addition, expression of *ICAM1* is also found to be higher in oral tongue squamous cell carcinoma as compared to normal tongue tissue [163]. *CXCL12*/*CXCR4* axis has been proposed to play a prominent role from early steps of oral malignant transformation to the progress of oral carcinogenesis [143]. Furthermore, *CXCL12/CXCR4* signaling in OSCC cells is also thought to be involved in invasion or micro-metastasis at the primary site and lymph node metastasis [152]. *PTK2* (also called focal adhesion kinase [FAK]) is a candidate gene that most likely drives the 8q24.3 amplification which in turn results in over-expression of *PTK2* mRNA and protein in OSCC cells [164]. Additionally, the expression level of focal adhesion kinase expression is known to be increased in invasive and pre-invasive oral cancers [155]. *COL1A2* is found to be among the list of genes that are over-expressed in oral carcinoma [165 166 167] including its over-expression in head and neck squamous cell carcinoma (HNSCC) [156]. Empirical studies that define the role of *ITGB1*in OSCC are scarce; however, it is known that a small non-coding micro-RNA down-regulate the expression of *ITGB1*, which in turn suppresses OSCC [139]. Somewhat similar to our approach but with different perspectives, Zhongyu et al. [167] performed a meta-analysis by integrated 4 public microarray OSCC datasets. Comparable to our findings, *COL1A2* was identified as one of the up-regulated genes in the OSCC tissues relative to controls. In a study performed by Bundela et al. involving two OSCC expression datasets [168], *PTK2* is found to be up-regulated. These genes were also associated with the processes that are well known to be involved in cancers such as cellular component movement, immune response, and cell adhesion (Additional file 2; Supplementary Table 3). Because empirical studies that represent the direct role of *ITGB1* and *PTK2* genes in oral cancer are limited, precise role of these genes in OSCC is not well established.

A pair-wise scatter plot for these 5 genes revealed that these were correlated, where healthy samples have lower expression values (red squares) compared to the tumor samples (blue circles) (Figure 3). Because our predicted anti-oral cancer genes could successfully distinguish tumor samples from normal controls, they are thought to be related to OSCC. Owing to the good correlation among these genes, classification models were developed using four popular state-of-the-art supervised classification methods to assess discriminative ability of candidate genes between healthy and tumor samples. The conditional inference trees, random forest and bagging ensemble, bagging, and SVM provided overall predictive accuracies (Q) of 83%, 87%, 84%, and 87%, respectively, with an average of 85.25%. Furthermore, AUCs of 0.84, 0.91, 0.87, and 0.89 were obtained for conditional inference trees, random forest and bagging ensemble, bagging, and SVM, respectively. As seen from the figure (Figure 4), all classifiers performed equally well during discrimination of tumor samples from the normal ones. The AUC values were beyond 0.50 in all classification methods; these results indicate that all of the classification algorithms performed better than random discrimination. The LOOCV accuracy of 85.80% also indicated an unbiased assessment and stability of genes selected. Taken all together, these observations indicate that the ensemble-based feature selection approach were able to capture more informative and compact set of genes which have the capability to discriminate between healthy and tumor samples; therefore, the identified five genes—*ICAM1*, *ITGB1*, *CXCR4*, *PTK2*, and *COL1A2*—were considered candidate "oral cancer genes".

**3.4. CXC chemokine receptor 4 (*CXCR4*) as an important drug target in oral cancer**

Considering the importance of central protein receptors, several attempts have been made to block them for therapeutic intervention [40 41] with an ultimate aim to interfere with downstream signaling molecules. Because *CXCR4* is both druggable [169 170] and highly relevant to oral cancer [143 144 171 153], there has been considerable interest in the clinical potential of *CXCR4* inhibitors [172]. However, *CXCR4* antagonists are reported with some side effects [173 174 175]; therefore, there is a strong need to discover novel molecules, of diverse structural and chemical features, with potential therapeutic value and lesser side effects. Several studies have assessed the potential anti-cancer properties of natural compounds [49 50 51 52], and these molecules provide specific scaffolds that make them comparable to trade drugs [176]. Overall, based on the known reports of *CXCR4* in oral cancer and effectiveness of natural compounds towards diseases, *CXCR4* was prospected for molecules of the plant origin.

### 3.5. Selection of optimal molecular descriptors of *CXCR4* inhibitors

Molecular feature selection methods are capable of improving the prediction accuracies and selection of meaningful features in cheminformatics research [58 61]. A total of 1,444 1D and 2D molecular descriptors—also referred to as features—were calculated for each *CXCR4* antagonists molecule. After removing redundant descriptors, performing feature selection, and considering those descriptors that presented significant difference in their values (p-value), the feature vectors space was reduced from 1,444 molecular descriptors to 10 significant ones. These 10 descriptors—maxHBint9, MPC9, minHBint5, VR3_Dt, SdssC, GATS6s, AATSC0c, VC-5, VR1_Dt, and SHBint7—can be broadly divided into five classes on the basis of their properties: atom type electro-topological state descriptors, path counts, detour matrix, chi cluster, and autocorrelation descriptors (Table 3). These descriptors are important for describing electro-topological state [177], connectivity-framework, and topological distances [178] of chemical compounds.

### 3.6. Construction of a descriptor-based classification model and screening of novel plant derived molecules

The RF and bagging ensemble method was employed to build a classifier model on the basis of 10 optimized molecular descriptors. During this procedure, we employed the same method as previously defined for gene classification problem (section 2.4) except that training and testing were carried out by 5-fold cross-validation procedure. The performance results of varying training-testing combinations did not present major differences in the performance parameters. The assessment parameters computed from each subset were averaged across all five subsets, achieving a final sensitivity, specificity, overall accuracy and MCC of 76%, 86%, 81%, and 0.63, respectively; this good predictive accuracy indicate that the selected molecular descriptors were able to discriminate *CXCR4* inhibitor from non-inhibitors. The technical details of these performance measures are provided in Supporting Information (Additional file 1: Supplementary Methods).

There are no hard and fast rules regarding the selection of best prediction model; however, the run that provides the highest prediction accuracy for training set may be considered. But, such kind of approach can be misleading because a model with the highest prediction accuracy may not necessarily produce the highest accuracy on the other test datasets due to overfitting. Therefore, for each of the five models, best combination of features was

selected by accessing their variable importance values by assessing the conditional variable importance. The variables are considered important if their importance value are above the absolute values of lowest negative-scoring variable; this assumption is based on the rationale that the importance of irrelevant variables varies randomly around zero [179]. In our study, an ensemble of all significant variables was considered to generate the final classification model which finally comprised of all ten molecular descriptors. Furthermore, to avoid any bias in the prediction, an independent validation set was also used for evaluation of the classification model. The validation dataset comprise of 21 inhibitors and 18 non-inhibitors categorized according to their IC50 values obtained from CHEMBL (https://www.ebi.ac.uk/chembl/) chemical structure database. We obtained 61% (12/21) sensitivity, 71% (12/18) specificity, 67% (24/39) accuracy, and a MCC value of 0.36 on the external data set; this result also indicate usefulness of a reasonably good classification model in the screening of unknown molecules as inhibitors of *CXCR4*.

The SerpentinaDB and Phytochemica databases are comprehensive repositories of molecules (with immense therapeutic properties) compiled by our group that are present in medicinal plants of the Himalayan mountain range. A total of 84, 574, 75, 56, 80 and 142 molecules were obtained from plants *Atropa belladonna*, *Catharanthus roseus*, *Heliotropium indicum*, *Picrorhiza kurroa*, *Podophyllum hexandrum*, and *Rauvolfia serpentina*, respectively, making a composite total of 1,011 molecules. The molecular descriptor-based classifier model was used to classify molecules and prioritize pharmacologically relevant molecules from a repository of PDMs; this was performed for probing of novel potential plant-based *CXCR4* inhibitors. The molecules that were present in more than the selected threshold of confidence score (>0.75) were considered as significant *CXCR4* inhibitors; this resulted in a hit list of 17 possible potent inhibitors (Table 4). Most of the hits obtained (~76%) are elected from *C. roseus* dataset indicating the potential of this plant to be of pharmacological relevance.

The similar property principle states that structurally similar molecules tend to have common biological properties [180]. All 17 hit molecules identified were assessed for structural diversity in order to group molecules with similar biological properties on the basis of similar scaffolds. To assess the diversity and unique molecular scaffolds (those were highly prevalent in identified hits), we used the following procedure: the set of molecules was first clustered in to their discrete groups—based on Tanimoto similarity measure [181]—and then we identified

maximum common substructures—based on subgraph enumeration and subgraph isomorphism testing [93]—in each of the respective cluster. The clustering results indicated that molecules were grouped together into three distinct clusters (Figure 5): cluster 1 (CARS0212, CARS0220, CARS0609, CARS0616 and CARS0617), cluster 2 (CARS0026, CARS0027, CARS0375, and CARS0385), and cluster 3 (CARS0610, HEIN0041 and CARS0465). This apparent separation between groups clearly indicates that compounds with similar scaffold were well clustered.

Furthermore, the common substructures analysis revealed the presence of varying structural scaffolds in each of the three independent clusters. Cluster 1 and 3 presented groups of molecules with isoprene scaffolds thus representing terpenes/terpenoids as a major chemical class; this molecular class has not largely been explored for designing of the *CXCR4* inhibitors. Cluster 2—the second largest cluster—contained scaffolds which comprised of six-membered rings with a nitrogen moiety thus representing indole alkaloids as the major class. Contrary to the terpenes/terpenoids, indole alkaloids and their derivatives currently represent one of the major class of *CXCR4* inhibitors [172]. A representative member of each of the cluster is shown in a figure (Figure 6). Furthermore, because the crystal structure of *CXCR4* is available [95], these identified molecules were also assessed for steric and electrostatic complementarity with the binding pocket.

**3.7. Screening through pharmacokinetic properties (ADME/T) and interaction studies**

In the direction of analyzing novel chemical entities, most of the drugs fails at early or late stages of drug discovery pipeline due to unwanted pharmacokinetics or toxicity problems[182 183]. In view of these, as a post-docking filter, the prioritized molecules were first assessed for *in-silico* ADME/T properties (Table 5) aimed at discarding those molecules which were either potentially toxic, exhibited poor ADME/T properties, or possessed non-drug like properties. By means of the Discovery studio package, the prediction of pharmacokinetic properties was performed using ADME/T descriptors, where descriptors perform prediction on the basis of chemical structure of the molecules. The module uses six mathematical models to quantitatively predict molecular properties by set of rules/keys summarized in supplementary data (Additional file 2: Supplementary Table 4). The following six ADME/T characteristics of molecules were considered in our studies: (i) ADMET solubility level (predicts the solubility of each compound in water), (ii) ADMET BBB (blood brain barrier penetration) level (predicts the BBB penetration of each compound), (iii) ADMET absorption level (predicts the absorption of compound ), (iv)

ADMET hepatotoxicity (predicts the occurrence of dose-dependent human hepatotoxicity), (v) ADMET CYP2D6 binding (predicts the cytochrome P450 2D6 enzyme inhibition), and (vi) plasma protein binding (PPB, predicts whether or not a compound is likely to be highly bound to blood carrier proteins). Additional information of these pharmacokinetic properties is presented in detail elsewhere [184]. Furthermore, the molecules were also was assessed for "drug-likeness" according to the default specifications of FAF-Drugs3. On the basis of above criteria, a total of eight drug-like molecules (Table 5) passed out the ADME/T filter and were subjected for subsequent binding studies with *CXCR4* receptor using molecular docking.

The accuracy of a docking algorithm is usually measured by the root mean square deviation (RMSD) between experimentally observed heavy atom positions of the ligand and those predicted by the algorithm, which is usually in the range of 1.5–2 Å [185]. Before performing actual interaction studies, docking protocol was first validated by comparing the native experimental conformation of the bound ligand (IT1t) in the crystal complex (PDB ID: 3OE6) with that of its computationally obtained binding conformation. For this, the coordinates of bound ligand IT1t were extracted from the complex and re-docked into the inhibitor binding site. The binding site comprises of residues Asp97**,** Cys186**,** and Glu288, which were reported to be involved in making critical inhibitory protein-ligand interactions in the IT1t-*CXCR4* complex [95].The backbone atom RMSD between the experimental conformation and best IT1t pose (on the basis of lowest docked energy/binding affinity) was 1.9 Å; this result confirms the quality of docking protocol and its suitability for predicting reliable binding modes of prioritized molecules.

Docking studies were carried out with eight pharmacologically viable molecules in order to find out their optimal conformations in the binding pocket of *CXCR4*. After molecular docking studies, a total of three lead molecules (Figure 7) were finally selected on the basis of interactions with critical residues (Asp97**,** Cys186, and Glu288) at the ligand binding site and best binding affinity values, where binding affinity is the sum of total intermolecular energy, total internal energy, and torsional free energy minus the energy of unbound system [98]. The predicted binding energy (kcal/mol) indicates the strength of ligand binding to the protein receptor, and the more negative is this energy, the stronger is the binding. Each of these 3 molecules, in their best binding poses, possessed binding affinities of -5.3, -7.2, and -5.1 for CARS610 (Linalool), CARS617 (α-Eudesmol), and HEIN0041 (beta-linalool), respectively.

Importantly, these molecules were making interactions (hydrogen bonding) with residues known to be critical for *CXCR4* inhibition (Figure 7). In addition, these molecules were also making hydrophobic contacts with other residues whose validity and exact role need to be validated empirically. Linalool is known to possess inhibitory effects against breast, colorectal, and liver cancer cells [186]; however, as far as our knowledge, its use in oral cancer research has not been well reported. A Pubmed (http://www.ncbi.nlm.nih.gov/pubmed) database-driven literature search for these compounds also ascertained that these lead molecules have not specifically been reported for the inhibition of *CXCR4* receptor; this confirms the novelty of these compounds and their opportunity to be used as novel *CXCR4* inhibitors for the treatment for OSCC and other human malignancies where precise role of *CXCR4* is described [187 188 189 190 169 191 192 193 194 195 196].

## 4. Conclusions

Major challenges in the cancer research include the prioritization of targets and also the identification of novel small-molecules that could inhibit the drug targets. Considering both aspects simultaneously, herein, an integrated computational pipeline has been established which links prioritized oral cancer drug targets with identification of small molecule inhibitors of plant origin. The systems-level approach that we have presented for OSCC may allow researchers to analyze large volumes of data and discover new potential drug targets in other complex human diseases. We expect that the three potential PDMs identified for *CXCR4* can be ideal for experimental studies as potential novel anti-oral cancer agents. The lead molecules reported herein may also provide better insights for designing potential *CXCR4* inhibitors with improved efficacy and fewer side effects.

## Abbreviations:

*CXCR4*, CXC chemokine receptor 4; DEGs, Differentially Expressed Genes; GO, Gene Ontology; IC, inhibitory concentration; OSCC, Oral Squamous Cell Carcinoma; PDMs, Plant-derived molecules; RF, Random forest

## Authors' contributions

VR and VA conceived and designed the study. VR conducted the data analysis. VA supervised and coordinated the study. Both authors read and approved the final version of the manuscript.

**Acknowledgments**

We thank CSIR-Institute of the Himalayan Bioresource Technology (CSIR-IHBT) project "Computational Systems and Network Biology" for the computational infrastructure. Infrastructural support in the form of the Bioinformatics Infrastructure Facility (BIF) was provided by the Department of Biotechnology, the Government of India. The authors also thank Shivalika Pathania for technical help in molecules structure preparation. VR was supported by a fellowship from the CSIR project "Physiological biochemical and molecular analysis of economically important plants for understanding and exploiting their growth, adaptation and metabolic mechanisms". This manuscript represents CSIR-IHBT communication number: 3889.

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References


1 W. N. William and V. A. Papadimitrakopoulou, *Cancer Prev. Res.*, 2013, **6**, 375–8.

2 C. J. Darby Weydert, B. B. Smith, L. Xu, K. C. Kregel, J. M. Ritchie, C. S. Davis and L. W. Oberley, *Free Radic. Biol. Med.*, 2003, **34**, 316–29.

3 W.-W. Lai, S.-C. Hsu, F.-S. Chueh, Y.-Y. Chen, J.-S. Yang, J.-P. Lin, J.-C. Lien, C.-H. Tsai and J.-G. Chung, *Anticancer Res.*, 2013, **33**, 1941–50.

4 E. Abdelfadil, Y.-H. Cheng, D.-T. Bau, W.-J. Ting, L.-M. Chen, H.-H. Hsu, Y.-M. Lin, R.-J. Chen, F.-J. Tsai, C.-H. Tsai and C.-Y. Huang, *Am. J. Chin. Med.*, 2013, **41**, 683–96.

5 C. A. Granville and P. A. Dennis, *Am. J. Respir. Cell Mol. Biol.*, 2005, **32**, 169–76.

6 H. Zhang, C.-Y. Yu and B. Singer, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 4168–72.

7 S. Polager and D. Ginsberg, *Nat. Rev. Cancer*, 2009, **9**, 738–48.

8 P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman and M. R. Stratton, *Nat. Rev. Cancer*, 2004, **4**, 177–83.

9 K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen and X.-M. Zhao, *BMC Bioinformatics*, 2012, **13**, 126.

10 X. Varelas, M. P. Bouchie and M. A. Kukuruzinska, *Glycobiology*, 2014, **24**, 579–91.

11 R. Lu, F. Markowetz, R. D. Unwin, J. T. Leek, E. M. Airoldi, B. D. MacArthur, A. Lachmann, R. Rozov, A. Ma'ayan, L. A. Boyer, O. G. Troyanskaya, A. D. Whetton and I. R. Lemischka, *Nature*, 2009, **462**, 358–62.

12 P. K. Kreeger and D. A. Lauffenburger, *Carcinogenesis*, 2010, **31**, 2–8.

13 B. Zhang and S. Horvath, *Stat. Appl. Genet. Mol. Biol.*, 2005, **4**, Article17.

14 A. E. Ivliev, P. A. C. 't Hoen and M. G. Sergeeva, *Cancer Res.*, 2010, **70**, 10060–70.

15 S. Horvath, A. N. M. Nazmul-Hossain, R. P. E. Pollard, F. G. M. Kroese, A. Vissink, C. G. M. Kallenberg, F. K. L. Spijkervet, H. Bootsma, S. A. Michie, S. U. Gorr, A. B. Peck, C. Cai, H. Zhou and D. T. W. Wong, *Arthritis Res. Ther.*, 2012, **14**, R238.

16 J. Zhang, Y. Xiang, L. Ding, K. Keen-Circle, T. B. Borlawsky, H. G. Ozer, R. Jin, P. Payne and K. Huang, *BMC Bioinformatics*, 2010, **11 Suppl 9**, S5.

17 L. Wang, H. Tang, V. Thayanithy, S. Subramanian, A. L. Oberg, J. M. Cunningham, J. R. Cerhan, C. J. Steer and S. N. Thibodeau, *Cancer Res.*, 2009, **69**, 9490–7.

18 H. N. Kadarmideen, N. S. Watson-Haigh and N. M. Andronicos, *Mol. Biosyst.*, 2011, **7**, 235–46.

19 A. Kommadath, H. Bao, A. S. Arantes, G. S. Plastow, C. K. Tuggle, S. M. D. Bearson, L. L. Guan and P. Stothard, *BMC Genomics*, 2014, **15**, 452.

20 W.-C. Chou, A.-L. Cheng, M. Brotto and C.-Y. Chuang, *BMC Genomics*, 2014, **15**, 300.

21 M. N. Patel, M. D. Halling-Brown, J. E. Tym, P. Workman and B. Al-Lazikani, *Nat. Rev. Drug Discov.*, 2013, **12**, 35–50.

22 G. L. Verdine and L. D. Walensky, *Clin. Cancer Res.*, 2007, **13**, 7264–70.

23 K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabási, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–90.

24 E.-Á. Horvát, J. D. Zhang, S. Uhlmann, Ö. Sahin and K. A. Zweig, *PLoS One*, 2013, **8**, e73413.

25 A.-L. Barabási, N. Gulbahce and J. Loscalzo, *Nat. Rev. Genet.*, 2011, **12**, 56–68.

26 S. Pathania and V. Acharya, *Plant Mol. Biol. Report.*, 2015.

27 E. E. Schadt, *Nature*, 2009, **461**, 218–23.

28 M. Meyerson, S. Gabriel and G. Getz, *Nat. Rev. Genet.*, 2010, **11**, 685–96.

29 C. S. H. Tan, B. Bodenmiller, A. Pasculescu, M. Jovanovic, M. O. Hengartner, C. Jørgensen, G. D. Bader, R. Aebersold, T. Pawson and R. Linding, *Sci. Signal.*, 2009, **2**, ra39.

30 L. C. Molina, L. Belanche and A. Nebot, in *2002 IEEE International Conference on Data Mining*, eds. V. Kumar, S. Tsurnoto, N. Zhong, P. S. Yu and X. Wu, Maebashi City, Japan, 2002, pp. 306 – 313.

31 H. Liu, Motoda and Hiroshi, *Feature selection for knowledge discovery and data mining*, Springer US, 1st edn., 1998.

32 Y. Lee and C.-K. Lee, *Bioinformatics*, 2003, **19**, 1132–1139.

33 T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science (80-. ).*, 1999, **286**, 531–7.

34 L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M.

Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, *Nature*, 2002, **415**, 530–6.

35 Q. Liu, A. H. Sung, Z. Chen, J. Liu, L. Chen, M. Qiao, Z. Wang, X. Huang and Y. Deng, *BMC Genomics*, 2011, **12 Suppl 5**, S1.

36 Z. Jagga and D. Gupta, *BMC Proc.*, 2014, **8**, S2.

37 A. C. Tan and D. Gilbert, *Appl. Bioinformatics*, 2003, **2**, S75–83.

38 J. A. Cruz and D. S. Wishart, *Cancer Inform.*, 2006, **2**, 59–77.

39 Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.-M. Ngai and J. Shao, *Mol. Biosyst.*, 2015, **11**, 791–800.

40 Y. Shen, J. Liu, G. Estiu, B. Isin, Y.-Y. Ahn, D.-S. Lee, A.-L. Barabási, V. Kapatral, O. Wiest and Z. N. Oltvai, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 1082–7.

41 G. L. Johnson, H. G. Dohlman and L. M. Graves, *Curr. Opin. Chem. Bol.*, 2005, **9**, 325–31.

42 T. C. Hung, K. B. Chen, H. J. Huang and C. Y. Chen, *Evidence-Based Complement. Altern. Med.*, 2014, **2014**, 13.

43 S. Pathania, V. Randhawa and G. Bagler, *PLoS One*, 2013, **8**, e61327.

44 D.-L. Ma, D. S.-H. Chan and C.-H. Leung, *Chem. Sci.*, 2011, **2**, 1656–1665.

45 L. Zhao and R. D. Brinton, *J. Med. Chem.*, 2005, **48**, 3463–3466.

46 J. Shen, X. Xu, F. Cheng, H. Liu, X. Luo, J. Shen, K. Chen, W. Zhao, X. Shen and H. Jiang, *Curr. Med. Chem.*, 2003, **10**, 2327–2342.

47 J. M. Rollinger, H. Stuppner and T. Langer, in *Natural Compounds as Drugs Volume I*, eds. F. Petersen and R. Amstutz, Birkhäuser Basel, Basel, 2008, vol. 65, pp. 211–249.

48 S. Suhitha, S. K. Devi, K. Gunasekaran, H. C. Pakyntein, A. Bhattacharjee and D. Velmurugan, *Curr. Top. Med. Chem.*, 2015, **15**, 21–36.

49 A. Tariq, S. Mussarat and M. Adnan, *J. Ethnopharmacol.*, 2015, **164**, 96–119.

50 G. M. Cragg and D. J. Newman, *J. Ethnopharmacol.*, 2005, **100**, 72–9.

51 J. de D. Tamokou, J. R. Chouna, E. Fischer-Fodor, G. Chereches, O. Barbos, G. Damian, D. Benedec, M. Duma, A. P. N. Efouet, H. K. Wabo, J. R. Kuiate, A. Mot and R. Silaghi-Dumitrescu, *PLoS One*, 2013, **8**, e55880.

52 V. Kuete, H. K. Wabo, K. O. Eyong, M. T. Feussi, B. Wiench, B. Krusche, P. Tane, G. N. Folefoc and T. Efferth, *PLoS One*, 2011, **6**, e21762.

53 E. C. Barnes, V. Choomuenwai, K. T. Andrews, R. J. Quinn and R. A. Davis, *Org. Biomol. Chem.*, 2012, **10**, 4015–23.

54 J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 468–481.

55 J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honório and A. B. F. da Silva, *Curr. Med. Chem.*, 2012, **19**, 4289–97.

56 Q. Zang, D. M. Rotroff and R. S. Judson, *J. Chem. Inf. Model.*, 2013, **53**, 3244–61.

57 Z. Zhao, G. Fu, S. Liu, K. M. Elokely, R. J. Doerksen, Y. Chen and D. E. Wilkins, *BMC Bioinformatics*, 2013, **14 Suppl 1**, S16.

58 Y. Xue, Z. R. Li, C. W. Yap, L. Z. Sun, X. Chen and Y. Z. Chen, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1630–8.

59 E. Hazai, I. Hazai, I. Ragueneau-Majlessi, S. P. Chung, Z. Bikadi and Q. Mao, *BMC Bioinformatics*, 2013, **14**, 130.

60 N. C. Karthikeyan Ramaswamy, Mohamed Sadiq, Sridhar V, *J. Proteomics Bioinform.*, 2009, **2**.

61 H. Singh, S. Singh, D. Singla, S. M. Agarwal and G. P. S. Raghava, *Biol. Direct*, 2015, **10**, 10.

62 S. Bhavani, A. Nagargadde, A. Thawani, V. Sridhar and N. Chandra, *J. Chem. Inf. Model.*, 2015, **46**, 2478–86.

63 Y. Li, L. Wang, Z. Liu, C. Li, J. Xu, Q. Gu and J. Xu, *Mol. Biosyst.*, 2015, **11**, 1241–50.

64 G.-Z. Li, H.-H. Meng, W.-C. Lu, J. Y. Yang and M. Q. Yang, *BMC Bioinformatics*, 2008, **9 Suppl 6**, S7.

65 S. Sengupta and S. Bandyopadhyay, *Int. J. Comput. Biol.*, 2014, 1, 56–62.

66 S. Smusz, R. Kurczab and A. J. Bojarski, *J. Cheminform.*, 2013, **5**, 17.

67 R. Siegel, D. Naishadham and A. Jemal, *CA. Cancer J. Clin.*, 2012, **62**, 10–29.

68 V. Jayaswal, S.-J. Schramm, G. J. Mann, M. R. Wilkins and Y. H. Yang, *BMC Res. Notes*, 2013, **6**, 430.

69 V. Randhawa and V. Acharya, *BMC Med. Genomics*, 2015, **8**, 39.

70 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc.*, 1995, **57**, 289–300.

71 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–504.

72 S. Mitra, S. Das, S. Das, S. Ghosal and J. Chakrabarti, *Oral Oncol.*, 2012, **48**, 117–9.

73 A. E. Levine and D. L. Steffen, *Nucleic Acids Res.*, 2001, **29**, 300–302.

74 N. S. Gadewal and S. M. Zingde, *Bioinformation*, 2011, **6**, 169–70.

75 S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal and M. R. Stratton, *Curr. Protoc. Hum. Genet.*, 2008, **Chapter 10**.

76 X. Liu, Z.-P. Liu, X.-M. Zhao and L. Chen, *J. Am. Med. Informatics Assoc.*, **19**, 241–8.

77 G. Yu, .

78 S. Falcon and R. Gentleman, *Bioinformatics*, 2007, **23**, 257–8.

79 E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock, *Bioinformatics*, 2004, **20**, 3710–5.

80 A. Liaw and M. Wiener, *R news*, 2002, **2**, 18–22.

81 H. Zou and T. Hastie, *J. R. Stat. Soc. Ser. B*, 2005, **67**, 301–320.

82 T. Hothorn, K. Hornik and A. Zeileis, *J. Comput. Graph. Stat.*, 2006, **15**, 651–674.

83 A. Peters and T. Hothorn, 2015.

84 D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch, 2014.

85 T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics*, 2005, **21**, 3940–1.

86 T. A. Halgren, *J. Comput. Chem.*, 1999, **20**, 720–729.

87 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–74.

88 T. Wei, 2013.

89 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582–94.

90 S. Pathania, S. M. Ramakrishnan, V. Randhawa and G. Bagler, *BMC Complement. Altern. Med.*, 2015, **15**, 262.

91 S. Pathania, S. M. Ramakrishnan and G. Bagler, *Database J. Biol. Databases Curation*, 2015.

92 Cao, Y, Charisi, A, Cheng, L. C, Jiang, T, Girke and T, *Bioinformatics*, 2008, **24**, 1733–1734.

93 Y. Cao, T. Jiang and T. Girke, *Bioinformatics*, 2008, **24**, i366–74.

94 D. Lagorce, O. Sperandio, J. B. Baell, M. A. Miteva and B. O. Villoutreix, *Nucleic Acids Res.*, 2015, **43**, W200–7.

95 B. Wu, E. Y. T. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov and R. C. Stevens, *Science (80-. ).*, 2010, **330**, 1066–71.

96 J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter and E. E. Abola, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 1998, **54**, 1078–1084.

97 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785–91.

98 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–61.

99 R. A. Laskowski and M. B. Swindells, *J. Chem. Inf. Model.*, 2011, **51**, 2778–2786.

100 C. E. Huggins, A. A. Domenighetti, M. E. Ritchie, N. Khalil, J. M. Favaloro, J. Proietto, G. K. Smyth, S. Pepe and L. M. D. Delbridge, *J. Mol. Cell. Cardiol.*, 2008, **44**, 270–80.

101 A. Raouf, Y. Zhao, K. To, J. Stingl, A. Delaney, M. Barbara, N. Iscove, S. Jones, S. McKinney, J. Emerman, S. Aparicio, M. Marra and C. Eaves, *Cell Stem Cell*, 2008, **3**, 109–18.

102 M. J. Peart, G. K. Smyth, R. K. van Laar, D. D. Bowtell, V. M. Richon, P. A. Marks, A. J. Holloway and R. W. Johnstone, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 3697–702.

103 K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, 1st edn., 2009.

104 E. Wang, A. Lenferink and M. O'Connor-McCourt, *Cell. Mol. Life Sci.*, 2007, **64**, 1752–62.

105 C.-Y. Lin, Y.-M. Jeng, H.-Y. Chou, H.-C. Hsu, R.-H. Yuan, C.-P. Chiang and M. Y.-P. Kuo, *J. Surg. Oncol.*, 2008, **97**, 544–50.

106 K. Misawa, T. Kanazawa, Y. Misawa, A. Imai, S. Endo, K. Hakamada and H. Mineta, *Cancer Biomarkers*, 2012, **10**, 135–44.

107 N. Kuribayashi, D. Uchida, M. Kinouchi, N. Takamaru, T. Tamatani, H. Nagai and Y. Miyamoto, *PLoS One*, 2013, **8**, e80773.

108 J.-Y. Chuang, P.-C. Chen, C.-W. Tsao, A.-C. Chang, M.-Y. Lein, C.-C. Lin, S.-W. Wang, C.-W. Lin and C.-H. Tang, *Oncotarget*, 2015, **6**, 4239–52.

109 Y. Jin, C. Wang, X. Liu, W. Mu, Z. Chen, D. Yu, A. Wang, Y. Dai and X. Zhou, *J. Biol. Chem.*, 2011, **286**, 40104–9.

110 J.-Y. Chuang, Y.-L. Huang, W.-L. Yen, I.-P. Chiang, M.-H. Tsai and C.-H. Tang, *Int. J. Mol. Sci.*, 2014, **15**, 545–559.

111 V. Brailo, V. Vucićević-Boras, A. Cekić-Arambasin, I. Z. Alajbeg, A. Milenović and J. Lukac, *Oral Oncol.*, 2006, **42**, 370–3.

112 A. Lotfi, G. Mohammadi, A. Tavassoli, M. Mousaviagdas, H. Chavoshi and L. Saniee, *Asian Pacific J. Cancer Prev.*, 2015, **16**, 1327–30.

113 E. Vairaktaris, C. Yapijakis, Z. Serefoglou, A. Vylliotis, J. Ries, E. Nkenke, J. Wiltfang, S. Derka, S. Vassiliou, I. Springer, P. Kessler and F. W. Neukam, *Oral Oncol.*, 2006, **42**, 888–92.

114 K. Laimer, G. Spizzo, P. Obrist, G. Gastl, T. Brunhuber, G. Schäfer, B. Norer, M. Rasse, M. C. Haffner and W. Doppler, *Cancer*, 2007, **110**, 326–333.

115 X. Han, Y. Han, H. Jiao and Y. Jie, *Mol. Cells*, 2015, **38**, 112–21.

116 N. K. Carneiro, J. M. M. Oda, R. Losi Guembarovski, G. Ramos, B. V Oliveira, I. J. Cavalli, E. M. de S F Ribeiro, M. S. B. Gonçalves and M. A. E. Watanabe, *Int. J. Immunogenet.*, 2013, **40**, 292–8.

117 F. Supek, M. Bošnjak, N. Škunca and T. Šmuc, *PLoS One*, 2011, **6**, e21800.

118 A. Schlicker, F. S. Domingues, J. Rahnenführer and T. Lengauer, *BMC Bioinformatics*, 2006, **7**, 302.

119 S. I. Grivennikov, F. R. Greten and M. Karin, *Cell*, 2010, **140**, 883–99.

120 S. K. Choudhari, M. Chaudhary, S. Bagde, A. R. Gadbail and V. Joshi, *World J. Surg. Oncol.*, 2013, **11**, 118.

121 E. T. Roussos, J. S. Condeelis and A. Patsialou, *Nat. Rev. Cancer*, 2011, **11**, 573–87.

122 S. B. Jakowlew, *Cancer Metastasis Rev.*, 2006, **25**, 435–57.

123 T. Okegawa, R.-C. Pong, Y. Li and J.-T. Hsieh, *Acta Biochim. Pol.*, 2004, **51**, 445–57.

124 L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein and P. D'Eustachio, *Nucleic Acids Res.*, 2009, **37**, D619–22.

125 M. Mali, H. Andtfolk, H. M. Miettinen and M. Jalkanen, *J. Biol. Chem.*, 1994, **269**, 27795–8.

126 M. W. Pickup, J. K. Mouw and V. M. Weaver, *EMBO Rep.*, 2014, **15**, 1243–53.

127 P. A. Netti, D. A. Berk, M. A. Swartz, A. J. Grodzinsky and R. K. Jain, *Cancer Res.*, 2000, **60**, 2497–503.

128 G. Bendas and L. Borsig, *Int. J. Cell Biol.*, 2012, **2012**.

129 C. Boccaccio and P. M. Comoglio, *Cancer Res.*, 2005, **65**, 8579–82.

130 S.-W. Chang, S. Abdul-Kareem, A. F. Merican and R. B. Zain, *BMC Bioinformatics*, 2013, **14**, 170.

131 Y. Zhang, C. Ding and T. Li, *BMC Genomics*, 2008, **9 Suppl 2**, S27.

132 X. Li, Y. Fujikura, Y. H. Wang, T. Sawada, N. Tokuda, R. S. Lovely, Y. Hayatsu, T. Fukumoto and F. Shinozaki, *J. Oral Pathol. Med.*, 1997, **26**, 371–6.

133 C. M. Liu, T. S. Sheen, J. Y. Ko and C. T. Shun, *Br. J. Cancer*, 1999, **79**, 360–2.

134 G. T. Huang, X. Zhang and N. H. Park, *Int. J. Oncol.*, 2000, **17**, 479–86.

135 K. Perschbacher, L. Jackson-Boeters and T. Daley, *J. Oral Pathol. Med.*, 2004, **33**, 230–6.

136 S. I. Maruya, J. N. Myers, R. S. Weber, D. I. Rosenthal, R. Lotan and A. K. El-Naggar, *Oral Oncol.*, 2005, **41**, 580–8.

137 K. Sundelin, K. Roberg, R. Grénman and L. Håkansson, *J. Oral Pathol. Med.*, 2007, **36**, 177–83.

138 S.-F. Yang, M.-K. Chen, Y.-S. Hsieh, T.-T. Chung, Y.-H. Hsieh, C.-W. Lin, J.-L. Su, M.-H. Tsai and C.-H. Tang, *J. Biol. Chem.*, 2010, **285**, 29808–16.

139 S. Hunt, A. V Jones, E. E. Hinsley, S. A. Whawell and D. W. Lambert, *FEBS Lett.*, 2011, **585**, 187–92.

140 P. Dillenburg-Pilla, V. Patel, C. M. Mikelis, C. R. Zárate-Bladés, C. L. Doçi, P. Amornphimoltham, Z. Wang, D. Martin, K. Leelahavanichkul, R. T. Dorsam, A.

Masedunskas, R. Weigert, A. A. Molinolo and J. S. Gutkind, *FASEB J.*, 2015, **29**, 1056–68.

141 T. Yu, K. Liu, Y. Wu, J. Fan, J. Chen, C. Li, Q. Yang and Z. Wang, *Oncogene*, 2014, **33**, 5017–27.

142 D. Uchida, N. Kuribayashi, M. Kinouchi, G. Ohe, T. Tamatani, H. Nagai and Y. Miyamoto, *Clin. Exp. Metastasis*, 2013, **30**, 133–42.

143 J. Xia, N. Chen, Y. Hong, X. Chen, X. Tao, B. Cheng and Y. Huang, *Mediators Inflamm.*, 2012, **2012**.

144 T. Yu, Y. Wu, J. I. Helman, Y. Wen, C. Wang and L. Li, *Mol. Cancer Res.*, 2011, **9**, 161–72.

145 D. Uchida, T. Onoue, N. Kuribayashi, Y. Tomizuka, T. Tamatani, H. Nagai and Y. Miyamoto, *Eur. J. Cancer*, 2011, **47**, 452–9.

146 H. H. Oliveira-Neto, E. T. Silva, C. R. Leles, E. F. Mendonça, R. de C. Alencar, T. A. Silva and A. C. Batista, *Tumour Biol.*, 2008, **29**, 262–71.

147 X. Meng, L. Wuyi, X. Yuhong and C. Xinming, *J. Oral Pathol. Med.*, 2010, **39**, 63–8.

148 D.-S. Wen, X.-L. Zhu, S.-M. Guan, Y.-M. Wu, L.-L. Yu and J.-Z. Wu, *Oral Oncol.*, 2008, **44**, 545–54.

149 M. Taki, K. Higashikawa, S. Yoneda, S. Ono, H. Shigeishi, M. Nagayama and N. Kamata, *Oncol. Rep.*, 2008, **19**, 993–8.

150 D. Uchida, N.-M. Begum, Y. Tomizuka, T. Bando, A. Almofti, H. Yoshida and M. Sato, *Lab. Investig.*, 2004, **84**, 1538–46.

151 T. Ishikawa, K.-I. Nakashiro, S. Hara, S. K. Klosek, C. Li, S. Shintani and H. Hamakawa, *Int. J. Oncol.*, 2006, **28**, 61–6.

152 A. Almofti, D. Uchida, N. M. Begum, Y. Tomizuka, H. Iga, H. Yoshida and M. Sato, *Int. J. Oncol.*, 2004, **25**, 65–71.

153 C. B. Delilbasi, M. Okura, S. Iida and M. Kogo, *Oral Oncol.*, 2004, **40**, 154–7.

154 J.-Y. Chuang, W.-Y. Yang, C.-H. Lai, C.-D. Lin, M.-H. Tsai and C.-H. Tang, *Int. Immunopharmacol.*, 2011, **11**, 948–54.

155 L. J. Kornberg, *Head Neck*, 1998, **20**, 634–9.

156 H. Ye, T. Yu, S. Temam, B. L. Ziober, J. Wang, J. L. Schwartz, L. Mao, D. T. Wong and X. Zhou, *BMC Genomics*, 2008, **9**, 69.

157 F. Kaleağasıoğlu and M. R. Berger, *Oncol. Rep.*, 2014, **31**, 1407–16.

158 C.-N. Yeh, W.-H. Weng, G. Lenka, L.-C. Tsao, K.-C. Chiang, S.-T. Pang, T.-W. Chen, Y.-Y. Jan and M.-F. Chen, *Mol. Med. Rep.*, 2013, **8**, 350–60.

159 X. Lü, W. Chen and C. Zhang, *J. South. Med. Univ.*, 2011, **31**, 1197–9.

160 M. J. Ravosa, J. Ning, Y. Liu and M. S. Stack, *Arch. Oral Biol.*, 2011, **56**, 491–8.

161 C.-J. Chiu, M.-L. Chang, C.-P. Chiang, L.-J. Hahn, L.-L. Hsieh and C.-J. Chen, *Cancer Epidemiol. Biomarkers Prev.*, 2002, **11**, 646–53.

162 Y. Usami, K. Ishida, S. Sato, M. Kishino, M. Kiryu, Y. Ogawa, M. Okura, Y. Fukuda and S. Toyosawa, *Int. J. Cancer*, 2013, **133**, 568–78.

163 J. Yan, Y. Jiang, M. Ye, W. Liu and L. Feng, *J. Cancer Res. Ther.*, 2014, **10 Suppl**, C125–30.

164 K. M. Quesnelle, A. M. Sparano, Y. Wang, G. S. Weinstein and M. S. Brose, *Cancer Res.*, 2006, **66**, 798.

165 P. P. Reis, L. Waldron, B. Perez-Ordonez, M. Pintilie, N. N. Galloni, Y. Xuan, N. K. Cervigne, G. C. Warner, A. A. Makitie, C. Simpson, D. Goldstein, D. Brown, R. Gilbert, P. Gullane, J. Irish, I. Jurisica and S. Kamel-Reid, *BMC Cancer*, 2011, **11**, 437.

166 C. Chen, E. Méndez, J. Houck, W. Fan, P. Lohavanichbutr, D. Doody, B. Yueh, N. D. Futran, M. Upton, D. G. Farwell, S. M. Schwartz and L. P. Zhao, *Cancer Epidemiol. Biomarkers Prev.*, 2008, **17**, 2152–62.

167 Z. Liu, Y. Niu, C. Li, Y. Yang and C. Gao, *Head Neck*, 2012, **34**, 1789–1797.

168 S. Bundela, A. Sharma and P. S. Bisen, *PLoS One*, 2014, **9**, e102610.

169 A. Müller, B. Homey, H. Soto, N. Ge, D. Catron, M. E. Buchanan, T. McClanahan, E. Murphy, W. Yuan, S. N. Wagner, J. L. Barrera, A. Mohar, E. Verástegui and A. Zlotnik, *Nature*, 2001, **410**, 50–6.

170 C. C. Bleul, M. Farzan, H. Choe, C. Parolin, I. Clark-Lewis, J. Sodroski and T. A. Springer, *Nature*, 1996, **382**, 829–33.

171 A. O. Rehman and C. Wang, *Int. J. Oral Sci.*, 2009, **1**, 105–18.

172 B. Debnath, S. Xu, F. Grande, A. Garofalo and N. Neamati, *Theranostics*, 2013, **3**, 47–75.

173 D. Mukherjee and J. Zhao, *Am. J. Cancer Res.*, 2013, **3**, 46–57.

174 J. A. Burger and A. Peled, *Leukemia*, 2009, **23**, 43–52.

175 T. Murakami, S. Kumakura, T. Yamazaki, R. Tanaka, M. Hamatake, K. Okuma, W. Huang, J. Toma, J. Komano, M. Yanaka, Y. Tanaka and N. Yamamoto, *Antimicrob. Agents Chemother.*, 2009, **53**, 2940–2948.

176 M. L. Lee and G. Schneider, *J. Comb. Chem.*, 2001, **3**, 284–9.

177 L. H. Hall and L. B. Kier, *J. Chem. Inf. Model.*, 1995, **35**, 1039–1045.

178 L. B. Kier and L. H. Hall, *Molecular connectivity in chemistry and drug research*, Academic Press, New York, 1976.

179 C. Strobl, J. Malley and G. Tutz, *Psychol. Methods*, 2009, **14**, 323–48.

180 A. L. Teixeira and A. O. Falcao, *J. Chem. Inf. Model.*, 2014, **54**, 1833–49.

181 G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2014, **57**, 3186–204.

182 K. Lentz, J. Raybon and M. W. Sinz, in *Drug Discovery: Practices, Processes, and Perspectives*, eds. J. J. Li and E. J. Corey, John Wiley & Sons, 2013, p. 570.

183 J. H. Lin and A. Y. Lu, *Pharmacol. Rev.*, 1997, **49**, 403–49.

184 P. Ponnan, S. Gupta, M. Chopra, R. Tandon, A. S. Baghel, G. Gupta, A. K. Prasad, R. C. Rastogi, M. Bose and H. G. Raj, *ISRN Struct. Biol.*, 2013, **2013**, 1–12.

185 D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat. Rev. Drug Discov.*, 2004, **3**, 935–49.

186 M.-Y. Chang and Y.-L. Shen, *Molecules*, 2014, **19**, 6694–706.

187 B. Furusato, A. Mohamed, M. Uhlén and J. S. Rhim, *Pathol. Int.*, 2010, **60**, 497–505.

188 J. Myers, Ed., *Oral Cancer Metastasis*, Springer New York, New York, NY, 2010.

189 J. A. Burger, M. Burger and T. J. Kipps, *Blood*, 1999, **94**, 3658–67.

190 R. Möhle, F. Bautz, S. Rafii, M. A. Moore, W. Brugger and L. Kanz, *Blood*, 1998, **91**, 4523–30.

191 Y. M. Li, Y. Pan, Y. Wei, X. Cheng, B. P. Zhou, M. Tan, X. Zhou, W. Xia, G. N. Hortobagyi, D. Yu and M.-C. Hung, *Cancer Cell*, 2004, **6**, 459–69.

192 M. Burger, A. Glodek, T. Hartmann, A. Schmitt-Gräff, L. E. Silberstein, N. Fujii, T. J. Kipps and J. A. Burger, *Oncogene*, 2003, **22**, 8093–101.

193 T. Kijima, G. Maulik, P. C. Ma, E. V Tibaldi, R. E. Turner, B. Rollins, M. Sattler, B. E. Johnson and R. Salgia, *Cancer Res.*, 2002, **62**, 6304–11.

194 J. Wang, J. Wang, Y. Sun, W. Song, J. E. Nor, C. Y. Wang and R. S. Taichman, *Cell. ignalling*, 2005, **17**, 1578–92.

195 M. V Barbolina, M. Kim, Y. Liu, J. Shepard, A. Belmadani, R. J. Miller, L. D. Shea and M. S. Stack, *Mol. Cancer Res.*, 2010, **8**, 653–64.

196 I. S. Zeelenberg, L. Ruuls-Van Stalle and E. Roos, *Cancer Res.*, 2003, **63**, 3833–9.

197 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Res.*, 2004, **32**, D449–51.

198 T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, *Nucleic Acids Res.*, 2009, **37**, D767–D772.

199 C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2011, **39**, D698–704.

200 S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. L. Brinkman, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. W. Hancock, R. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stümpflen, M. Tyers, P. Uetz, I. Xenarios and H. Hermjakob, *Nat. Methods*, 2012, **9**, 345–50.

201 A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen, *Nucleic Acids Res.*, 2013, **41**, D808–15.

202 H. Tamamura, T. Araki, S. Ueda, Z. Wang, S. Oishi, A. Esaka, J. O. Trent, H. Nakashima, N. Yamamoto, S. C. Peiper, A. Otaka and N. Fujii, *J. Med. Chem.*, 2005, **48**, 3280–9.

203 V. I. Pérez-Nueno, D. W. Ritchie, O. Rabal, R. Pascual, J. I. Borrell and J. Teixidó, *J. Chem. Inf. Model.*, 2008, **48**, 509–33.

204 S. Rusconi, M. Lo Cicero, O. Viganò, F. Sirianni, E. Bulgheroni, S. Ferramosca, A. Bencini, A. Bianchi, L. Ruiz, C. Cabrera, J. Martinez-Picado, C. T. Supuran and M. Galli, *Molecules*, 2009, **14**, 1927–37.

205 R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, Wiley-VCH, 2nd edn., 2009.

**Figure Legends**

Figure 1. Principal component analysis (PCA) plot. PCA plot indicating grouping of healthy and tumor samples by their expression profiles into two separate clusters. The X- and Y-axis are defined by the first and second principal components, respectively.

Figure 2. Network module visualization. The figure indicating disruption in correlation of gene expression among hub genes and their interaction partners in healthy (A) and cancerous (B) networks as indicated by the changes in edge color. Both networks are visualized using Cytoscape software package and the color scale ranges from red (strong negative correlation) through yellow (no correlation) to blue (strong positive correlation).

Figure 3. Classification of normal and tumor samples. A pair-wise scatter plot for identified five candidate oral cancer genes—*ICAM1*, *ITGB1*, *CXCR4*, *PTK2*, and *COL1A2*— revealing correlation among them, with normal tissues having lower expression values (red squares) than cancerous tissues (blue circles).

Figure 4. Receiver operating characteristic curve (ROC) plot of oral cancer genes. The ROC plot obtained for four classifiers viz., conditional inference trees, random forest and bagging ensemble, bagging, and support vector machine. ROC depicts True Positive Rate (sensitivity) versus False Positive Rate (1-specificity). The diagonal line in the ROC curve has an area under the curve (AUC) value of 0.5, representing the predictive power of a random guess. A smooth curve through a set of data points was obtained with locally weighted scatterplot smoothing (LOWESS) non-parametric regression method.

Figure 5. Multi-dimensional scaling (MDS) plot of identified hit molecules. MDS plot constructed on the basis of Tanimoto index showing a distinct discrimination of molecules into individual clusters where green, black and gray color represent cluster 1 (terpenes [CARS0212, CARS0220, CARS0609, CARS0616 and CARS0617]), cluster 2 (indole alkaloids [CARS0026, CARS0027, CARS0375, and CARS0385]) and cluster 3 (terpenes [CARS0610, HEIN0041 and CARS0465]), respectively.

Figure 6. Representative structures of molecules present in each individual cluster. (A) α-Bisabolol (CARS0212; cluster 1), (B) N, N-dimethyltryptamine (CARS0026; cluster 2), and (C) 1,8-Cineole (CARS0610; cluster 3).

Figure 7. The best plant-derived molecules (PDMs) as potent *CXCR4* inhibitors. 2D structures of three lead PDMs identified on the basis of binding affinity and interactions with critical residues: (A) CARS610 (Linalool), (B) CARS617 (α-Eudesmol), and (C) HEIN0041 (beta-linalool). Lower panel of figure shows the hydrogen bonds between hydroxyl group of lead molecules and *CXCR4* protein residues which are indicated in cyan color, while residues making hydrophobic contacts are represented in surface view.

**Tables**

Table 1. A list of databases, along with release date/version and number of protein-protein interactions (PPIs), used in the present study.

| Database Name | Release date/Version | No. of Human PPI | Reference |
|---|---|---|---|
| DIP | Jan 17, 2014 | 3,100 | [197] |
| HPRD | Apr 13, 2010/Release 9 | 35,348 | [198] |
| BIOGRID | April 1, 2014/3.2.111 | 49,958 | [199] |
| IMEx Consortium | Apr 10, 2014 | 58,250 | [200] |
| STRING | Dec 27, 2013/9.1 | 1,93,734 | [201] |

Table 2. The families of CXCR4 inhibitor molecules compiled for the present study.

| Family | Number of compounds | References |
|---|---|---|
| Cyclic peptides | 4 | 202 |
| Tetrahydroquinoline derivatives | 110 | 172 |
| Indole derivatives | 7 | 172 |
| AMD derivatives | 15 | 203 |
| Macrocyclic polyamines | 4 | 204 |

Table 3. The optimal molecular descriptors selected for *CXCR4* centred classification model building.

| Descriptor type | Descriptor | Reference |
|---|---|---|
| Atom type electrotopological state descriptors | maxHBint9; minHBint5; SdssC; SHBint7 | 177 |
| Autocorrelation | GATS6s; AATSC0c | 205 |
| Detour matrix | VR1_Dt; VR3_Dt | 205 |
| Path counts | MPC9 | 205 |
| Chi cluster | VC-5 | 178 |

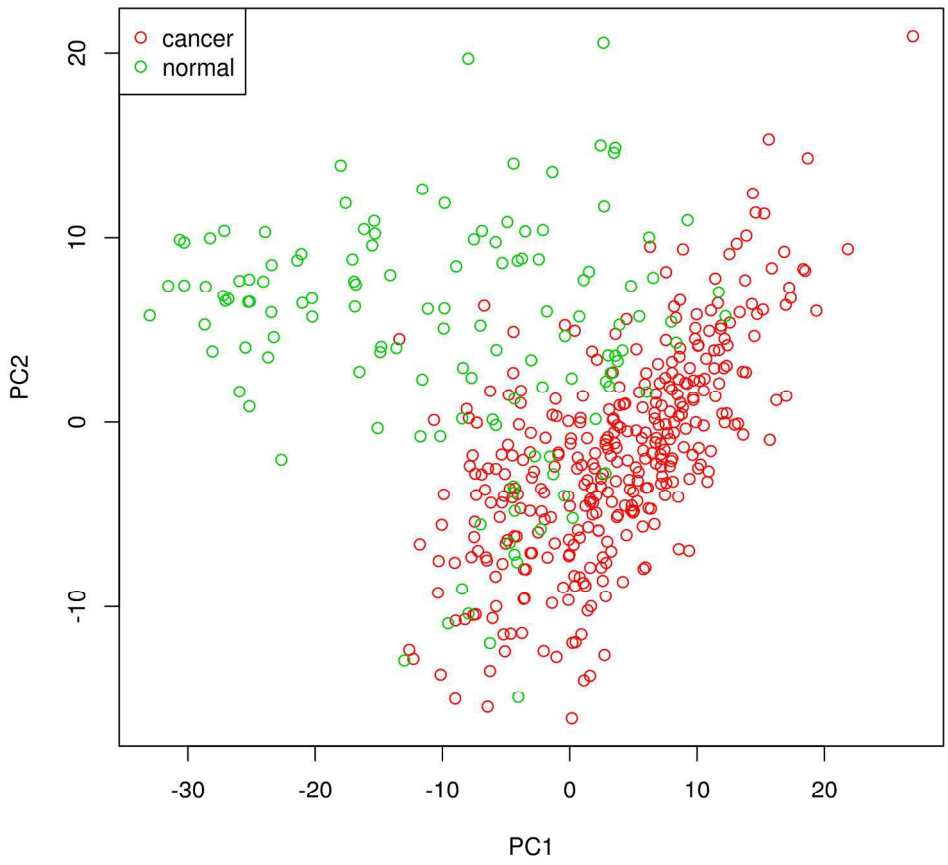
Table 4. A summary of hit compounds identified form screening/prioritization of plant derived molecules (PDMs).

| Molecule ID | Confidence Score | Plant | Chemical Name | Chemical Class |
|---|---|---|---|---|
| CARS0575 | 0.91 | *C. roseus* | Secodine | Indole alkaloid |
| CARS0026 | 0.89 | *C. roseus* | N, N-dimethyltryptamine | Indole alkaloid |
| CARS0027 | 0.86 | *C. roseus* | $N_b$-acetyltryptamine | Indole alkaloid |
| RASE0127 | 0.78 | *R. serpentina* | Suaveoline | Indole alkaloid |
| CARS0212 | 0.76 | *C. roseus* | α-Bisabolol | Monocyclic sesquiterpene alcohol |
| CARS0616 | 0.76 | *C. roseus* | γ-Eudesmol | Sesquiterpenoids |
| CARS0609 | 0.76 | *C. roseus* | 1,8-Cineole | Monoterpenoids |
| CARS0220 | 0.76 | *C. roseus* | Manool | Diterpenes |
| CARS0617 | 0.76 | *C. roseus* | α-Eudesmol | Sesquiterpenoids |
| CARS0376 | 0.76 | *C. roseus* | N-[(S)-α-Methylbenzyl]tetrahydro-γ-carboline | Indole alkaloid |
| CARS0375 | 0.76 | *C. roseus* | N-[(S)-α-Methylbenzyl]-4-piperidone | Indole alkaloid |
| HEIN0045 | 0.76 | *H. indicum* | Beta-Ionone | Terpenes |
| CARS0465 | 0.75 | *C. roseus* | (−)-piperitone | Monoterpenes |
| CARS0385 | 0.75 | *C. roseus* | N-[(S)-1-( l-Naphthyl)ethyl]-4-piperidone | Indole alkaloid |
| ATBE0034 | 0.75 | *A. belladonna* | N-methyl pyrroline | Heterocyclic alkaloids |
| HEIN0041 | 0.75 | *H. indicum* | beta-linalool | Terpenes |
| CARS0610 | 0.75 | *C. roseus* | Linalool | Terpene alcohols |

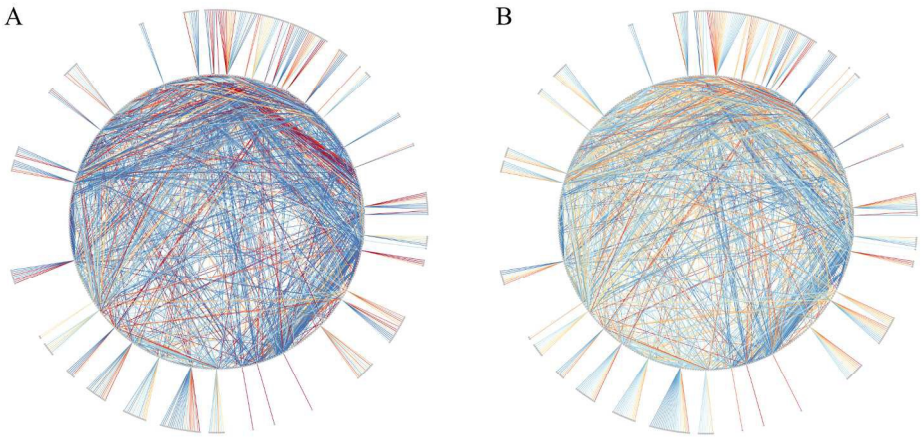
Table 5. The *in-silico* ADME/T properties of prioritized plant derived molecules (PDMs) obtained from descriptor-based classification model. The numerical values indicate the pharmacokinetic properties, and the highlighted rows indicate the molecule selected for docking studies (NA indicates not available)

| Molecule ID | A* | B* | C* | D* | E* | F* | G* | H* | I* | J* | K* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ATBE0034 | 4 | 4 | 2 | 0 | 1 | 0 | 83.13 | 0.49 | 0 | 0 | 1 |
| CARS0026 | NA | NA | NA | 0 | 1 | 0 | 188.27 | 2.35 | 3 | 1 | 2 |
| CARS0027 | NA | NA | NA | 0 | 1 | 0 | 202.25 | 1.62 | 3 | 2 | 3 |
| **CARS0212** | **2** | **0** | **0** | **0** | **0** | **1** | **222.37** | **3.79** | **4** | **1** | **1** |
| **CARS0220** | **2** | **0** | **0** | **0** | **0** | **1** | **290.48** | **5.72** | **4** | **1** | **1** |
| CARS0375 | NA | NA | NA | 1 | 0 | 0 | 203.28 | 1.61 | 2 | 0 | 2 |
| CARS0376 | NA | NA | NA | 1 | 1 | 0 | 276.38 | 3.85 | 2 | 1 | 2 |
| CARS0385 | NA | NA | NA | 1 | 1 | 1 | 253.34 | 2.86 | 2 | 0 | 2 |
| **CARS0465** | **3** | **1** | **0** | **0** | **0** | **1** | **152.23** | **2.18** | **1** | **0** | **1** |
| **CARS0575** | **NA** | **NA** | **NA** | **1** | **0** | **1** | **338.44** | **3.84** | **7** | **1** | **4** |
| CARS0609 | 3 | 1 | 0 | 0 | 1 | 1 | 154.25 | 2.18 | 0 | 0 | 1 |
| **CARS0610** | **3** | **1** | **0** | **0** | **0** | **1** | **154.25** | **2.73** | **4** | **1** | **1** |
| CARS0616 | 2 | 0 | 0 | 0 | 1 | 1 | 222.37 | 3.39 | 1 | 1 | 1 |
| **CARS0617** | **2** | **1** | **0** | **0** | **0** | **1** | **222.37** | **3.5** | **1** | **1** | **1** |
| **HEIN0041** | **3** | **1** | **0** | **0** | **0** | **1** | **154.25** | **3.21** | **4** | **1** | **1** |
| **HEIN0045** | **2** | **1** | **0** | **0** | **0** | **1** | **192.3** | **2.91** | **2** | **0** | **1** |
| RASE0127 | NA | NA | NA | 1 | 1 | 1 | 303.4 | 2.76 | 1 | 1 | 3 |

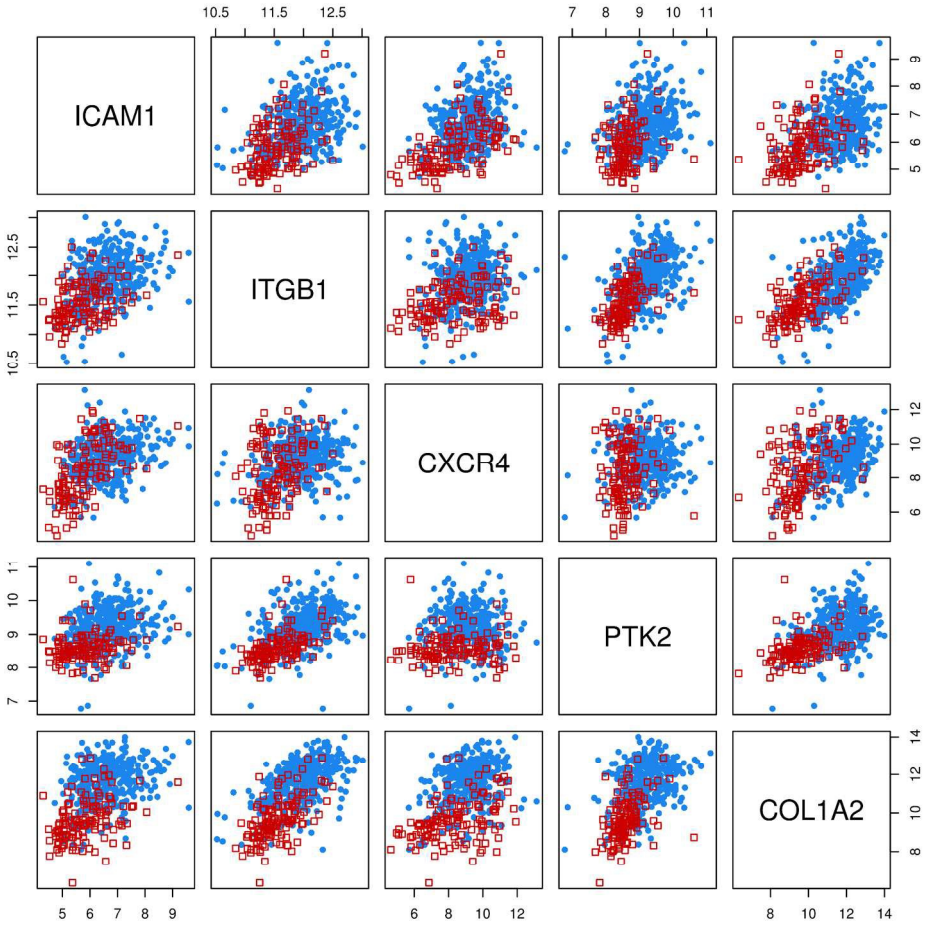
A*-K* represent ADMET solubility level, ADMET blood brain barrier (BBB) level, ADMET absorption level, ADMET CYP2D6 (predicted class), ADMET hepatotoxicity (predicted class), ADMET plasma protein binding level (PPB; predicted class), molecular weight, logP, rotatable bonds, hydrogen bond donors (HBD), and hydrogen bond acceptors (HBA), respectively.



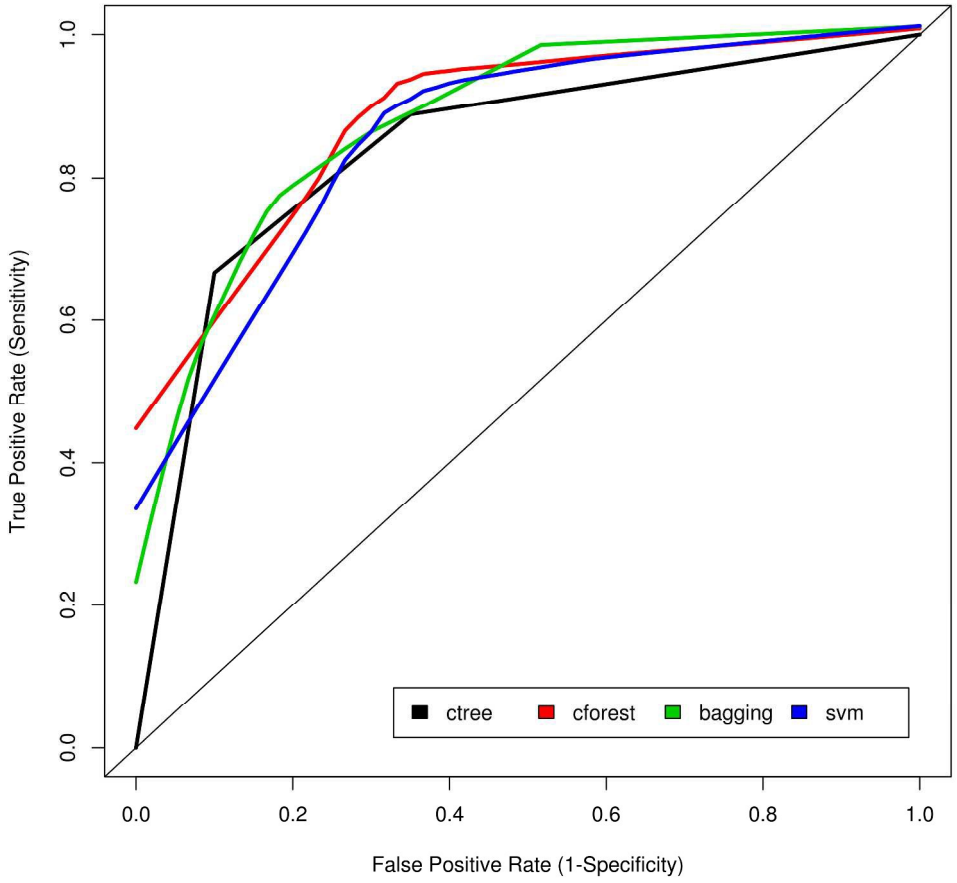

Principal component analysis (PCA) plot. PCA plot indicating grouping of healthy and tumor samples by their expression profiles into two separate clusters. The X- and Y-axis are defined by the first and second principal components, respectively.



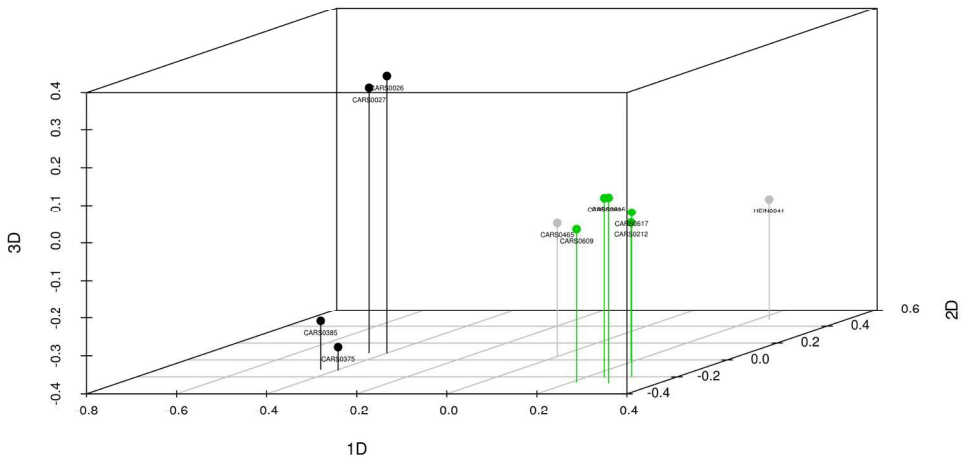

Network module visualization. The figure indicating disruption in correlation of gene expression among hub genes and their interaction partners in healthy (A) and cancerous (B) networks as indicated by the changes in edge color. Both networks are visualized using Cytoscape software package and the color scale ranges from red (strong negative correlation) through yellow (no correlation) to blue (strong positive correlation).



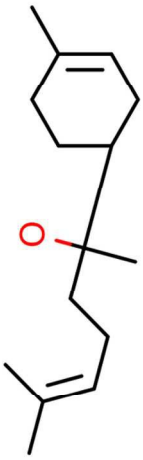

Classification of normal and tumor samples. A pair-wise scatter plot for identified five candidate oral cancer genes—ICAM1, ITGB1, CXCR4, PTK2, and COL1A2— revealing correlation among them, with normal tissues having lower expression values (red squares) than cancerous tissues (blue circles).



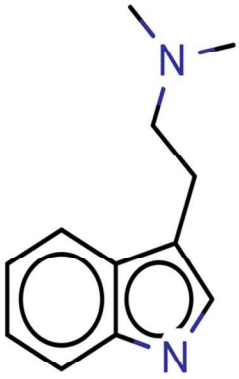

Receiver operating characteristic curve (ROC) plot of oral cancer genes. The ROC plot obtained for four classifiers viz., conditional inference trees, random forest and bagging ensemble, bagging, and support vector machine. ROC depicts True Positive Rate (sensitivity) versus False Positive Rate (1-specificity). The diagonal line in the ROC curve has an area under the curve (AUC) value of 0.5, representing the predictive power of a random guess. A smooth curve through a set of data points was obtained with locally weighted scatterplot smoothing (LOWESS) non-parametric regression method.



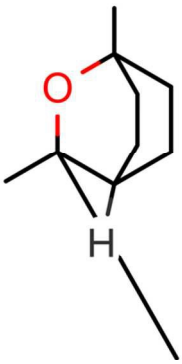

Multi-dimensional scaling (MDS) plot of identified hit molecules. MDS plot constructed on the basis of Tanimoto index showing a distinct discrimination of molecules into individual clusters where green, black and gray color represent cluster 1 (terpenes [CARS0212, CARS0220, CARS0609, CARS0616 and CARS0617]), cluster 2 (indole alkaloids [CARS0026, CARS0027, CARS0375, and CARS0385]) and cluster 3 (terpenes [CARS0610, HEIN0041 and CARS0465]), respectively.




α-Bisabolol

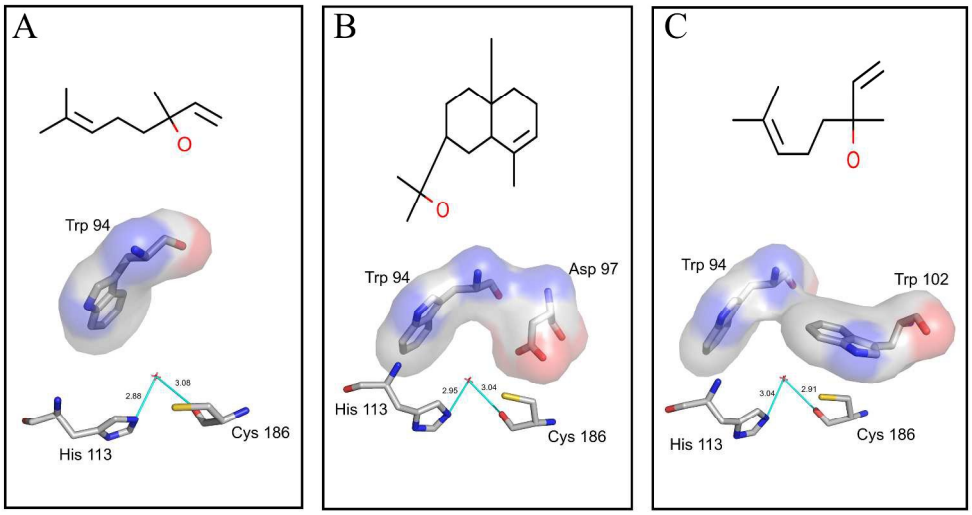



N, N-dimethyltryptamine




1,8-Cineole

Representative structures of molecules present in each individual cluster. (A) α-Bisabolol (CARS0212; cluster 1), (B) N, N-dimethyltryptamine (CARS0026; cluster 2), and (C) 1,8-Cineole (CARS0610; cluster 3).




The best plant-derived molecules (PDMs) as potent CXCR4 inhibitors. 2D structures of three lead PDMs identified on the basis of binding affinity and interactions with critical residues: (A) CARS610 (Linalool), (B) CARS617 (α-Eudesmol), and (C) HEIN0041 (beta-linalool). Lower panel of figure shows the hydrogen bonds between hydroxyl group of lead molecules and CXCR4 protein residues which are indicated in cyan color, while residues making hydrophobic contacts are represented in surface view.
881x473mm (96 x 96 DPI)