Cureus Part of Springer Nature

> Received 04/02/2024 Review began 04/22/2024 Review ended 05/03/2024 Published 05/08/2024

© Copyright 2024

Arumuganainar. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Extra Tree Classifier Predicts an Interactome Hub Gene as HSPB1 in Oral Cancer: A Bioinformatics Analysis

Deepavalli Arumuganainar¹

1. Department of Periodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, IND

Corresponding author: Deepavalli Arumuganainar, deepavallia.sdc@saveetha.com

Abstract

Introduction

Oral cancer is a significant global health issue that is mainly caused by factors, such as smoking, alcohol consumption, poor oral hygiene, age, and the human papillomavirus. Unfortunately, delayed diagnosis contributes to high rates of illness and mortality. However, saliva shows promise as a potential source for early detection, prognosis, and treatment. By analyzing the proteins and their interactions in saliva, we can gain insights that can assist in early detection and prediction. In this study, we aim to identify and predict the key genes, known as hub genes, in the salivary transcriptomics data of oral cancer patients and healthy individuals.

Methods

The data used for the analysis were obtained from salivaryproteome.org (https://salivaryproteome.org/) . The retrieved data consisted of individuals with oral cancer who were assigned unique identification numbers (IDs) 1025, 1030, 1027, and 1029, while the healthy individuals were assigned IDs 4256, 4257, 4255, and 4258, respectively. Differential gene expression analysis was used to identify genes that showed significant differences between the two groups. Uniformity and clustering were assessed through heatmaps and principal component analysis. Protein-protein interactions were investigated using the STRING database and Cytoscape. In addition, machine learning algorithms were employed to identify key genes involved in the interatomic interactions by analyzing transcriptomics data generated from the differential gene expression analysis.

Results

The accuracy and class accuracy of the extra tree classifier showed 98% and 97% in predicting interactomic hub genes, and HSPB1 was identified as a hub gene using Cytohubba from Cytoscape.

Conclusion

The predictive extra tree classifier, with its high accuracy in analysing interactomic hub genes in oral cancer, can potentially improve diagnosis and treatment strategies.

Categories: Genetics, Dentistry, Oncology Keywords: interactome analysis, hspb1, hub genes, oral cancer, interactome

Introduction

Oral cancer, a type of head and neck cancer, is a perilous and potentially life-threatening condition affecting the oral cavity, including the lips, tongue, gingiva, cheeks, and roof of the mouth [1,2]. It can be caused by smoking, alcohol consumption, a family history of cancer, and exposure to chemicals. Symptoms include persistent mouth sores, swelling, difficulty chewing, changes in voice or speech, and unexplained weight loss. Treatment options include surgical therapy, chemotherapy, radiation therapy, targeted therapy, or a combination of the above. Prevention involves avoiding risk factors, regular dental check-ups, and the HPV vaccine [3]. Advancements in medical technology and awareness have improved diagnosis and treatment outcomes. In addition, craniofacial tumors are known to present diverse clinical behaviors and histopathological presentations. A previous clinicopathological study of craniofacial tumor management was conducted on 319 patients and identified that the unique nature of cystic and neoplastic pathosis in the craniofacial region was due to aesthetic and functional derangements, with variations in prevalence due to occupational, sociocultural, and climatic factors [4].

The oral cancer interactome hub refers to a network of interactions between proteins involved in the development and progression of oral cancer. An interactome is a comprehensive map of interactions

How to cite this article

Arumuganainar D (May 08, 2024) Extra Tree Classifier Predicts an Interactome Hub Gene as HSPB1 in Oral Cancer: A Bioinformatics Analysis. Cureus 16(5): e59863. DOI 10.7759/cureus.59863

between proteins within a cell or biological system. The oral cancer interactome hub [5-7] is a research concept that aims to understand molecular pathways and protein-protein interactions in oral cancer. It uses experimental techniques and computational methods to map interactions, identify biomarkers, and develop new treatment strategies [8,9]. This knowledge can lead to targeted therapies that improve treatment outcomes for patients.

A previous study proposes a protein-protein interaction network technique to analyze genes involved in oral cancer disorders, revealing a network with 208 nodes and 1572 edges. TP53, a key gene [10,11], is identified as a key player in the oral cancer network, enhancing disease understanding and treatment. A study analyzing saliva proteomics data for oral cancer identified 74 candidate genes or proteins. The analysis identified nine hub-bottleneck proteins, including kininogen-1, angiotensinogen, annexin A1, IL-8, IgG heavy chains, CRP, collagen alpha-1, and fibronectin, as potential biomarkers for diagnosis and treatment. A study uses high-throughput transcriptomics methods and machine learning to identify disease biomarkers for MetS. The integrated approach uses WGCNA, LASSO regression, and RF algorithms for feature selection. A logistic regression model and web nomogram calculator are created for MetS risk assessment, aiming to develop reliable biomarkers [12]. Machine learning is used to predict and analyze complex interactions within biological networks, particularly in interactome hub genes. These genes play crucial roles in cellular processes and are often linked to disease mechanisms. Techniques like clustering, classification, and network analysis help identify key hub genes, predict gene function, discover disease biomarkers, and understand genetic perturbations.

Oral cancer results in high mortality rates and limited treatment options. This study aimed to identify interactomic hub genes associated with oral cancer and propose a predictive model for their identification by using differential expression analysis to identify genes significantly dysregulated in oral cancer and extract protein-protein interaction data from publicly available databases. Using functional enrichment analysis, the study will validate the identified hub genes' significance and functional relevance. A predictive model will be developed to classify oral cancer patients into subtypes or predict disease outcomes. Hence, our study is based on machine learning prediction of interactome hub genes in oral cancer.

Materials And Methods

The data used for the analysis were obtained from salivaryproteome.org. The retrieved data consisted of individuals with oral cancer who were assigned unique identification numbers (IDs) 1025, 1030, 1027, and 1029, while the healthy individuals were assigned IDs 4256, 4257, 4255, and 4258, respectively.

This study acknowledges the importance of integrated bioinformatics analysis in identifying hub genes using network analysis. This approach allows us to leverage the power of AI algorithms to predict hub genes, leading to the discovery of potential biomarkers and drug targets in oral cancer. The study aims to contribute to the field by shedding light on the intricacies of oral cancer at a molecular level. By utilizing network analysis and applying advanced AI algorithms, the key genes that play a crucial role in the development and progression of oral cancer may be uncovered. Thus, the findings may significantly enhance our understanding of oral cancer pathogenesis and potentially pave the way for developing novel therapeutic strategies.

Differentially expressed gene (DEG) analysis was performed to identify genes that exhibited significant changes in expression between the oral squamous cell carcinoma (OSCC) group and the healthy group. A heatmap was generated using the ggplot2 package to assess the uniformity of the samples. In addition, clustering analysis was performed using principal component analysis (PCA) to explore the overall gene expression patterns between the OSCC and healthy groups. DEG analysis was carried out using the limma package in R (version 4.1.0). Linear models were constructed to identify DEGs based on their expression levels. Genes with an absolute log2 (fold change) greater than 1.0 and an adjusted p-value of <0.05 were considered significant DEGs.

A Venn diagram was generated to identify critical genes common to the cancer and healthy groups and display the overlap between the samples. This allowed for the identification of genes that may have a potential role in the development or progression of oral cancer. To investigate protein-protein interactions (PPIs), the STRING database was utilized to construct a PPI network. This network provides insights into the interactions and functions of proteins associated with oral cancer. Subsequently, the PPI network was visualized using Cytoscape, a powerful network visualization and analysis tool. Furthermore, the cytoHubba plugin for Cytoscape was employed to identify the top 10 hub genes using the maximal clique centrality algorithm.

Extra tree classifier

The extra tree classifier is an ensemble-based machine learning algorithm, a variant of the decision tree algorithm. It consists of many individual decision trees trained independently and in parallel. Each tree is trained with random feature selection, threshold selection, and majority voting. After training, each tree independently predicts the target class based on input features. The class label with the most votes is selected as the predicted class label. This architecture introduces randomness to reduce overfitting and



increase the diversity of individual decision trees, leading to improved model accuracy and generalization.

Results

The results show a heat map of cancer and normal sample distribution. The color scale represents the gene expression levels, with blue indicating low expression, white indicating medium expression, and red indicating high expression. The "annotation_col" argument assigns a different color or annotation to the cancer and normal samples to distinguish them in the heatmap (Figure 1).



FIGURE 1: Heat map of cancer and normal sample distribution

The volcano plot provides a snapshot of changes in gene expression levels and their statistical significance. The plot uses color to differentiate genes based on their regulation status, with upregulated genes marked in red and downregulated genes in blue. The position and color of each gene point provide insights into its significance and direction of regulation. The interpretation of a volcano plot should consider the significance threshold, fold change cutoff values, and potential sources of noise and biases in the data (Figure 2).



FIGURE 2: Volcano plot providing a snapshot of changes in gene expression levels and their statistical significance

The PCA plot obtained in the study shows the distribution of samples or genes in a dataset, focusing on two principal components: PC-1 accounts for 51.1% of the total variation, capturing main patterns, while PC-2 explains 12.1%. The plot helps identify grouping or clustering patterns, as distant samples contribute the most to the observed variation. The percentage values indicate the variability each component explains, allowing researchers to identify relationships or groupings among samples or genes (Figure 3).



FIGURE 3: Principal component analysis (PCA) plot showing the distribution of samples or genes in a dataset, focusing on two principal components



The present study also provides an insightful examination of the protein interactome associated with the differential gene expression hubs. Furthermore, a meticulous analysis was conducted using M code written in Power Query language to identify smaller subsets of cohesive subgroups within larger clusters. Notably, applying CytoHubba analysis successfully identified a distinct sub-cluster among the differentially expressed genes in the hub and top hub genes, namely, SERPINC1, TTR, HSPB1, HP, and APOA1 (Figure 4).



FIGURE 4: Interactome of differential gene expression in oral cancer

The results of the extra tree classifier model for oral cancer classification indicate high performance in terms of accurately predicting the disease status (Table 1). The AUC value of 0.98 suggests that the model has a high discriminatory power in distinguishing between positive and negative cases of oral cancer. A higher AUC value indicates better model performance, with a value of 1.0 representing a perfect classifier. The CA of 0.97 indicates that the model correctly classified 97% of the instances in the test dataset. This high accuracy suggests that the model effectively differentiates between oral and non-cancer cases. The F1 score of 0.97 demonstrates the balance between precision and recall, with a higher score suggesting better performance. This indicates that the model can accurately classify both positive and negative cases of oral cancer), it is correct approximately 94% of the time. This metric indicates the model's accuracy in correctly identifying true positive samples and minimizing false positive predictions. Overall, the results suggest that the extra tree classifier model exhibits high accuracy, precision, and F1 score, indicating its effectiveness in predicting the interactomic hub genes of oral cancer. It demonstrates strong potential for clinical application in guiding decision-making and personalized treatment strategies.

Model	AUC	CA	F1	Precision
Extra tree classifier	0.98	0.97	0.97	0.94

TABLE 1: Extra tree classifier model performance with an accuracy score of 98%

The confusion matrix obtained from the study classifies the model's performance into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). A true positive rate of nearly 100% indicates that the model correctly identified most positive cases, while a true negative rate of nearly 100% indicates that the model accurately classified most negative cases. In conclusion, a nearly 100% true positive and true negative rate in the confusion matrix indicates that the model performs excellently in accurately predicting hub genes with high sensitivity and specificity (Figure 5).





FIGURE 5: Confusion matrix of true positives and true negatives of nearly 100%

Discussion

Saliva, a potential alternative to blood serum and urine for health monitoring, is gaining attention due to its affordability, safety, simplicity, and non-invasiveness. Saliva analysis offers several advantages, including non-invasive sample collection, easy storage, active patient participation, cost-effectiveness, and reduced risk of infection. "Salivaomics" involves the study of various aspects of saliva, such as its genomes, transcriptomics, proteomics, metabolomics, microbiomics, and microRNA. Using techniques like mass spectrometry and proteomics, researchers have identified over 3000 distinct proteins in saliva [11,13-15]. This wealth of data holds promises for customizing therapeutic approaches in the future. The diagnosis of oral cancer relies on comprehensive oral examinations and tissue biopsies for accurate detection. The progression of oral carcinogenesis involves various histopathologic stages and molecular changes, underscoring the importance of proteomics in understanding alterations in gene expression. In specific studies [16,17], researchers analyzed hub genes from saliva samples of individuals with oral cancer and compared them to those from healthy individuals. Identifying these hub genes provides important insights into the molecular pathways involved in oral cancer and can potentially improve diagnostic accuracy and treatment strategies (Figures 1-4).

SERPINC1, a serine protease inhibitor, regulates blood clotting by inhibiting coagulation factors. Mutations in the SERPINC1 gene can lead to inherited thrombophilia. TTR, a carrier protein, binds to thyroid hormones and vitamin A and prevents amyloid aggregation. Mutations in the TTR gene can cause amyloidosis. HSPB1, a small heat shock protein, aids in protein folding and cell cycle regulation. HP, a glycoprotein, binds hemoglobin to prevent oxidative damage and iron loss. Genetic variations in HP can influence susceptibility to certain diseases. APOA1, a major high-density lipoprotein (HDL) component, plays a crucial role in cholesterol metabolism. Mutations in the APOA1 gene can lead to disorders like familial hypoalphalipoproteinemia [1,18,19].

The study found that HSPB1 [16,17] expression in oral squamous carcinoma cells is weak compared to normal human oral keratinocytes, linked to promoter hypermethylation. Treatment with RG108 could induce HSPB1 expression, as seen in primary oral squamous carcinomas. High levels of HspB1 expression in cancer cells contribute to resistance and aggressiveness, leading to poor clinical outcomes in various cancers. HSPB1 is involved in tumor invasion, the formation of metastatic colonies, and resistance to cancer treatments. It targets client proteins to promote resistance to cell death and malignant behavior. Elevated levels of HSPB1 have been found in cancer stem cells, which play a role in their maintenance and induce long-term dormancy. HSPB1 also acts as a sensor of genetic imbalances and enhances cancer cells' resistance to various anti-cancer drugs. Multiplex network analysis identified 46 hub genes for oral cancer, including PIK3CG, PIK3R5, MYH7, CDC20, and CCL4, with high prediction accuracy. These genes may improve understanding of tumorigenesis and molecular events, opening new research routes for multi-omics biological data analysis [20].

One recent study identified 46 hub genes with 96% prediction accuracy, particularly PIK3CG, PIK3R5, MYH7, CDC20, and CCL4, which have significant biological implications for oral cancer and offered new research routes [21]. A study on head and neck squamous cell carcinoma patients identified 65 concordant genes, including a 13-gene panel. Validation in the Oncomine database revealed significant over-expression of all 13 genes, with six genes (CXCL8, CXCL10, FN1, PLAU, SERPINE1, and SPP1) significantly associated with



prognosis. This panel could improve diagnostic, prognostic, and treatment approaches [22]. Similar to these previous study results, the current study has identified HSPB1 as one of the interactome hub genes.

The predictive extra tree classifier has shown high accuracy in analyzing interactomic hub genes in oral cancer, with a 98% accuracy rate (Figure 5). This suggests that the identified genes are crucial in oral cancer development and progression.

Limitations

However, the model needs validation in larger, more diverse cohorts to ensure generalizability. In addition, the study may have a potential bias due to prioritizing genes with higher connectivity. However, these prioritizing or hub genes could be better identified by additional in-silico computational approaches.

Future directions include integrating other omics data, developing non-invasive saliva collection methods, and standardizing saliva collection and analysis protocols. Further research is needed to fully understand the potential of interactomic hub genes in oral cancer diagnosis and treatment.

Conclusions

The predictive extra tree classifier's high accuracy in analyzing interactomic hub genes in oral cancer suggests its potential for improving diagnosis and treatment strategies. However, validation in larger cohorts and integration of omics data are needed for generalizability and robustness. Future research should focus on non-invasive saliva collection methods and larger cohorts.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Deepavalli Arumuganainar

Acquisition, analysis, or interpretation of data: Deepavalli Arumuganainar

Drafting of the manuscript: Deepavalli Arumuganainar

Critical review of the manuscript for important intellectual content: Deepavalli Arumuganainar

Supervision: Deepavalli Arumuganainar

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue. Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: Payment/services info: All authors have declared that no financial support was received from any organization for the submitted work. Financial relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. Other relationships: All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

- Feller LL, Khammissa RR, Kramer BB, Lemmer JJ: Oral squamous cell carcinoma in relation to field precancerisation: pathobiology. Cancer Cell Int. 2013, 13:31. 10.1186/1475-2867-13-31
- Comprehensive genomic characterization of head and neck squamous cell carcinomas . Nature. 2015, 517:576-82. 10.1038/nature14129
- Forner D, Noel CW, Wu V, et al.: Nonsurgical management of resectable oral cavity cancer in the wake of COVID-19: a rapid review and meta-analysis. Oral Oncol. 2020, 109:104849. 10.1016/j.oraloncology.2020.104849
- 4. Afrah A Aldelaimi, Hamid H Enezei, Tahrir N Aldelaimi, Khalil Abdulla Mohammed.: Tumors of Craniofacial Region in Iraq (Clinicopathological Study). J Res Med Dent Sci. 2021, 9 (1):66-71.
- Yadalam PK, Anegundi RV, Ramadoss R, Joseph B, Veeramuthu A: Felodipine repurposed for targeting TRPV1 receptor to relieve oral cancer pain. Oral Oncol. 2022, 134:106094. 10.1016/j.oraloncology.2022.106094
- Patil S, Yadalam PK, Hosmani J, Khan ZA, Ahmed ZH, Shankar VG, Awan KH: Oral immune-mediated disorders with malignant potential/association: an overview. Dis Mon. 2023, 69:101349. 10.1016/j.disamonth.2022.101349
- 7. Patil S, Yadalam PK, Hosmani J, et al.: Modulation of oral cancer and periodontitis using chemotherapeutic

agents - a narrative review. Dis Mon. 2023, 69:101348. 10.1016/j.disamonth.2022.101348

- Yadalam PK, Krishnamurthi I, Srimathi R, et al.: Gene and protein interaction network analysis in the epithelial-mesenchymal transition of Hertwig's Epithelial Root Sheath reveals periodontal regenerative drug targets - an in silico study. Saudi J Biol Sci. 2022, 29:3822-9. 10.1016/j.sjbs.2022.03.007
- 9. Yadalam PK, Sivasankari T, Rengaraj S, et al.: Gene interaction network analysis reveals IFI44L as a drug target in rheumatoid arthritis and periodontitis. Molecules. 2022, 27:10.3390/molecules27092749
- Purwaningsih NM, Khor GH, Nik Mohd Rosdy NM, Abdul Rahman EO: Wnt pathway in oral cancer: a review update. Saudi Dent J. 2021, 33:813-8. 10.1016/j.sdentj.2021.08.002
- 11. Markopoulos AK, Michailidou EZ, Tzimagiorgis G: Salivary markers for oral cancer detection. Open Dent J. 2010, 4:172-8. 10.2174/1874210601004010172
- 12. Leemans CR, Snijders PJ, Brakenhoff RH: The molecular landscape of head and neck cancer . Nat Rev Cancer. 2018, 18:269-82. 10.1038/nrc.2018.11
- McCravy MS, Ingram JL: A Hint from Wnt: squamous cell differentiation in the airways . Am J Respir Cell Mol Biol. 2023, 68:601-2. 10.1165/rcmb.2023-0065ED
- 14. Zeng Y, Kong Y, Liao L, Zhu H: Involvement of RUVBL1 in WNT/β-catenin signaling in oral squamous cell carcinoma. Dis Markers. 2022, 2022:3398492. 10.1155/2022/3398492
- Li Q, Luo H, Dai FQ, et al.: SAMD9 promotes postoperative recurrence of esophageal squamous cell carcinoma by stimulating myh9-mediated GSK3β/β-catenin signaling. Adv Sci (Weinh). 2023, 10:e2203573. 10.1002/advs.202203573
- Amiri Dash Atan N, Koushki M, Rezaei Tavirani M, Ahmadi NA: Protein-protein interaction network analysis of salivary proteomic data in oral cancer cases. Asian Pac J Cancer Prev. 2018, 19:1639-45. 10.22034/APJCP.2018.19.6.1639
- 17. Arrigo AP, Gibert B: HspB1, HspB5 and HspB4 in human cancers: potent oncogenic role of some of their client proteins. Cancers (Basel). 2014, 6:333-65. 10.3390/cancers6010333
- Whitmore SE, Lamont RJ: Oral bacteria and cancer. PLoS Pathog. 2014, 10:e1003933. 10.1371/journal.ppat.1003933
- Balkwill F, Mantovani A: Inflammation and cancer: back to Virchow?. Lancet. 2001, 357:539-45. 10.1016/S0140-6736(00)04046-0
- Sulfianti A, Karimah N, Nurhasanah A: In silico analysis of HLA-1 and HLA-2 recognition of a designed recombinant human papillomavirus vaccine based on L1 protein HPV subtype 45. J Genet Eng Biotechnol. 2023, 21:167. 10.1186/s43141-023-00593-8
- Mahapatra S, Bhuyan R, Das J, Swarnkar T: Integrated multiplex network based approach for hub gene identification in oral cancer. Heliyon. 2021, 7:e07418. 10.1016/j.heliyon.2021.e07418
- Thakore VP, Patel KD, Bhadresha KP, Patel PS, Jain NK: An integrative analysis to enumerate candidate genes for clinical use in oral cancer. J Cancer Res Ther. 2022, 18:S182-90. 10.4103/jcrt.JCRT_1607_20