Identification of Thyroid Carcinoma Related Genes with mRMR and Shortest Path Approaches

Yaping Xu*, Yue Deng*, Zhenhua Ji, Haibin Liu, Yueyang Liu, Hu Peng, Jian Wu*, Jingping Fan*

Department of Otolaryngology head and neck surgery, Shanghai Changzheng Hospital, Second Military Medical University, Shanghai, China

Abstract

Thyroid cancer is a malignant neoplasm originated from thyroid cells. It can be classified into papillary carcinomas (PTCs) and anaplastic carcinomas (ATCs). Although ATCs are in an very aggressive status and cause more death than PTCs, their difference is poorly understood at molecular level. In this study, we focus on the transcriptome difference among PTCs, ATCs and normal tissue from a published dataset including 45 normal tissues, 49 PTCs and 11 ATCs, by applying a machine learning method, maximum relevance minimum redundancy, and identified 9 genes (BCL2, MRPS31, ID4, RASAL2, DLG2, MY018, ZBTB8, PRKCG and PPP6C) as important candidates involved in the progression of thyroid cancer. We further identified the protein-protein interaction (PPI) sub network from the shortest paths among the 9 genes in a PPI network constructed based on STRING database. Our results may provide insights to the molecular mechanism of the progression of thyroid cancer.

Citation: Xu Y, Deng Y, Ji Z, Liu H, Liu Y, et al. (2014) Identification of Thyroid Carcinoma Related Genes with mRMR and Shortest Path Approaches. PLoS ONE 9(4): e94022. doi:10.1371/journal.pone.0094022

Editor: Shoba Ranganathan, Macquarie University, Australia

Received October 22, 2013; Accepted March 12, 2014; Published April 9, 2014

Copyright: © 2014 Xu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by Foundation for Military Medicine, China (BWS111035). The founders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jianwu2008@126.com (JW); fanjingp@163.com (JF)

† These authors contributed equally to this work.

Introduction

Thyroid tumors include encapsulated benign tumors and carcinomas, and carcinomas can be classified into papillary carcinomas (PTCs) and anaplastic carcinomas (ATCs). Although frequency of ATC is low (<5%), it is in a very aggressive status of thyroid carcinomas, responsible for about half of its death and its patients have a short survival time after diagnosis (6 month in average) [1]. ATC is evolved from PTC, and they are found to share genetic alterations [2]. However, limited studies reported their difference at transcriptome level [2–5], resulting a lack of systematic analysis of its tumor evolution.

In order to bring insight into the progression of thyroid carcinomas at systems level, we adopted a two-step computational strategy [6]. By using an effective machine learning method—mRMR (maximum relevance, minimum redundancy), we first identify genes responsible for the progressing transcriptome difference among normal tissue, PTC and ATC using the mRNA microarray data from Hebrant et al.’s study [5]. The machine learning method mRMR does not only identify genes with independent effect along, but also take the redundancy effect among genes selected into account. Additional to the pipeline used by Li et al. [6], we applied different validation methods, such as leave-one-out validation, 10 fold cross validation and stratified 10 fold cross validation, to determine the number of genes which separate the three tissue status, due to one validation method along may provide biased information of prediction accuracy of the machine learning model. Second, we address the function of these genes at systems level by integrating known protein-protein interaction (PPI) from STRING database. A network of shortest paths among the genes from a background PPI network could be further revealed.

Materials and Methods

Transcriptome Array Dataset

We adopted the gene expression data of thyroid cancer from Hebrant et al.’s study [5], which include the transcriptome array data of 11 anaplastic thyroid carcinomas (ATCs), 49 papillary thyroid carcinomas (PTCs) and 45 normal thyroids (Normal) based on Affymetrix Human Genome U133 Plus 2.0 Array. This dataset was retrieved from NCBI Gene Expression Omnibus (GEO) with an accession number GSE33630. The array platform is with 54,675 probes corresponding to 20,283 protein coding genes. The array signals were normalized with RMA using the Affymetrix Bioconductor package. For the expression value of a gene, we used the average value of normalized signals of its corresponding probes.

STRING PPI data

The PPI data was retrieved from STRING database (version 9.0) (http://string.embl.de/) [7]. The PPI data includes both known and predicted protein interactions. We constructed a PPI network based on the STRING PPI data using a R package ‘igraph’ [8]. In the network, proteins are presented as nodes of the networks and edges corresponding to the protein-protein interactions.
The mRMR algorithm

We used mRMR (maximum relevance minimum redundancy) method to define a gene set which can separate the three sample sets (ATC, PTC and Normal). The mRMR was first used in analyzing microarray data by Peng et al. [9]. Its idea is to rank features according to their relevance to the target sample variable, and meanwhile take redundancy among the features into consideration. So genes in the selected gene set has the best trade-off between maximum relevance to phenotype and minimum redundancy within genes in the selected set.

Using mutual information (MI) defined using equation (1), we quantified relevance as well as redundancy,

$$I(x,y) = \int p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) dx dy$$

where $p(x,y)$ is a joint probabilistic density of vectors $x$ and $y$, and $p(x)$ and $p(y)$ are marginal probabilistic densities.

Relevance $D$ between a gene $f$ and its target variable $c$ is defined as,

$$D = I(f,c)$$

And redundancy $R$ between gene $f$ and genes in gene set $T$ is defined as,

$$R = \frac{1}{m} \sum_{g_i \in T} I(f,g_i)$$

where $m$ is the number of genes in $T$. The trade-off between relevance and redundancy is obtained as follows,

**Table 1. The 10 Genes selected using mRMR and IFS.**

| Gene Name | Entrez Gene ID | mRMR score |
|-----------|----------------|-------------|
| BCL2      | 596            | 1.09662945  |
| MRPS31    | 10240          | 0.222372096 |
| ID4       | 3400           | 0.32164204  |
| RASAL2    | 9462           | 0.39051335  |
| DLG2      | 1740           | 0.33428422  |
| MY01B     | 4430           | 0.35448678  |
| ZBTB5     | 9925           | 0.38452316  |
| LOC646736 |                | 0.33957166  |
| PRKCC1    | 5588           | 0.35941048  |
| PPP6C     | 5537           | 0.34092868  |

doi:10.1371/journal.pone.0094022.t001

Figure 1. IFS curve of the classification of ATCs, PTCs and normal tissue samples. The X-axis indicate the number of genes used for classification/prediction, and Y-axis is the prediction accuracies by NNA evaluated using leave-one-out (orange), 10 fold (green) and stratified 10 fold (blue) cross validation.

doi:10.1371/journal.pone.0094022.g001
max_{f}(D - R)

(4)

Repeating the above calculation a gene set is selected to distinguish target variables under mRMR condition with a given number \(N\) of genes.

Using incremental feature selection (IFS), the number \(N\) can be determined. Its idea is to compare prediction accuracy defined in the following selection among different \(N\)s, and choose the one with highest accuracy.

**Prediction of phenotypes**

We used the widely used Nearest Neighbor Algorithm (NNA) to predict the target variable [10]. “Nearness” is calculated as follows,

\[ N(x_1, x_2) = 1 - \frac{x_1 \cdot x_2}{|x_1||x_2|} \]

(5)

where \(x_1\) and \(x_2\) are two vectors of genes representing two samples. The smaller \(N(x_1, x_2)\) is, the more similar the two samples are [11,12].

**Model Validation**

In Li et al.'s study [6], leave-one-out validation was applied to validate the prediction accuracy of the study. Although the advantages of this validation method is explain in some studies [6,13], we noticed that there are other theoretical studies demonstrated there are bias in the estimation of accuracy in the leave-one-out validation in many circumstances [14,15]. In order to provide more information of the accuracy of the prediction model and to give an accurate estimation of the number of genes separate different tumor status, we applied two additional validation methods – 10 fold cross validation [14] and stratified 10 fold cross validation because of the stratification of tumor status (normal, PTC and ATC) [15].

**Shortest paths tracing**

Genes do not function only by itself, but also by its interaction with others as well as environmental factors. Protein-protein interaction (PPI) network would bring us insights into the comprehensive biological systems. We attempted to provide such insights by searching the shortest paths which link the genes selected using mRMR and IFS in PPI network constructed according to STRING PPI data. The shortest paths were estimated using Dijkstra’s algorithm [16].

**Enrichment analysis**

GO (Gene Ontology) term enrichment and KEGG pathway enrichment were performed using DAVID tools [17]. We estimated the \(P\) values, corrected \(P\) values with Benjamin multiple testing correction which controlled family-wide false discovery rate, and fold enrichment values for each functional or pathway terms.

**Results**

Ten candidate genes identified by mRMR, NNA and IFS

On the basis of mRMR estimation, we tested the predictor of NNA described in the Materials and Methods section, with one feature, two features, … to 400 features. The result of IFS curve representing prediction accuracy estimated by leave-one-out, 10 fold and stratified 10 fold cross validation, compared with the number of features is shown in Figure 1. We noticed that although the estimation accuracies different among the three different methods, but the minimum number of genes required separating tumor status is approximately the same – about 9 or 10 (Figure 1 and Table S1). We selected 10 genes to include more candidates for further analysis and studies, and the accuracy was 0.848, 0.857 and 0.877 for leave-one-out, 10 fold and stratified 10 fold cross validation.
validation separately. The top 10 genes selected using mRMR include 9 known genes (BCL2, MRPS31, ID4, RASAL2, DLG2, MY01B, ZBTB5, PRKCQ, PPP6C), and a miscRNA (miscellaneous RNA, LOC646736) (Table 1). Interestingly, the 10 candidate genes have no overlap with the 9 differentially expression gene between ATC and PTC identified in the Hebrant et al.’s study. One of the possible reasons is that in our detection, we considered the variation in transcriptome differences in normal tissue, ATC and PTC together.

Table 3. KEGG pathway enrichment of the 25 genes selected on the shortest paths.

| Term                        | Gene Count | P Value     | Fold Enrichment |
|-----------------------------|------------|-------------|-----------------|
| T cell receptor signaling pathway | 4          | 0.002282004 | 13.45238095     |
| Neurotrophin signaling pathway | 4          | 0.003385035 | 11.71658986     |
| Pathways in cancer          | 5          | 0.007626354 | 5.536803136     |
| Small cell lung cancer      | 3          | 0.018690317 | 12.97193878     |
| Apoptosis                   | 3          | 0.019971101 | 12.52463054     |
| Prostate cancer             | 3          | 0.02084525  | 12.24317817     |
| Thyroid cancer              | 2          | 0.071736805 | 25.04926108     |

Figure 2. 17 shortest paths genes among the 9 genes identified with mRMR methods. We identified 17 genes located on the shortest paths of STRING PPI network among the 9 mRMR identified genes. Genes in blue are those identified with mRMR methods, and genes in red are located on their shortest paths. The network is constructed based on STRING PPI data. doi:10.1371/journal.pone.0094022.g002
Shortest path genes
We constructed an undirected network based on PPI data from STRING using ‘igraph’ [8]. Then we traced shortest path between each pair of two genes from the 9 candidate genes identified using mRMR, in the PPI network using Dijkstra’s algorithm [16]. There are 16 genes located on the shortest path among the 9 candidate genes, and we presented them according to their network betweenness in the shortest paths composed sub-PPI network (Table 2 and Figure 2).

Enrichment of the 9 candidate genes and shortest paths genes
Using DAVID tools [17], we analyzed the functional enrichment of the 9 candidate genes together with 16 shortest path genes in KEGG pathway and GO term separately. For KEGG enrichment, the 25 genes are enriched in 7 KEGG pathways listed with their P value and fold enrichment value in Table 3. Interestingly, we found most of these pathways are important pathways related with cancer, such as T cell receptor signaling pathway, apoptosis, pathways in cancer, small cell lung cancer, prostate cancer, and thyroid cancer. T Cell Receptor (TCR) activation promotes several important signals that determine cell fate through regulating cytokine production, cell survival, proliferation, and differentiation. And T cells are especially important in cell-mediated immunity, which is the defense against tumor cells. More detailed functions of TCR in cancer is reviewed in Reference [18]. Moreover, thyroid cancer pathway was also found enriched by the set of the 25 genes. For GO term enrichment, 262 GO terms are enriched (Table S2). Several of them are related with cancer progression, like GO:0042127 regulation of cell proliferation, GO:0042980 regulation of apoptosis and GO:0043067 regulation of programmed cell death. These results provide circumstantial evidence supporting our data analysis pipeline.

Discussion

Genes identified by mRMR and IFS
We identified 9 genes, BCL2, MRPS31, ID4, RASAL2, DLG2, MY01B, ZBTB5, PRKCQ and PPP6C, and a miscRNA LOC646736 related with thyroid carcinoma in this study. Many of them are previously known important genes with thyroid development or progression, additional to the genes identified previously [5]. We further revealed the PPI network of the proteins coded by these genes by estimating the shortest path of the interactions based on a background PPI network constructed based on STRING database. Our results may provide important insights to understand the mechanism of the thyroid cancer progression at transcriptome level.

Supporting Information

Table S1 Prediction accuracy of three validation methods. (XLSX)
Table S2 GO enrichment of the 25 genes on the shortest paths. (TXT)

Acknowledgments

Thanks to Hongqiang Wu (Biotree Bio-technology Co., Ltd., Shanghai, China) for providing helps in data analysis, and we also very appreciate the useful comments and suggestions from the reviewers.

Author Contributions

Conceived and designed the experiments: JF. Performed the experiments: JW. Analyzed the data: YD ZJ HL YL HP. Wrote the paper: YX YD JW.
References

1. Ain KB (1999) Anaplastic thyroid carcinoma: A therapeutic challenge. Seminars in Surgical Oncology 16: 64–69.

2. Smallridge RC, Marlow LA, Copland JA (2009) Anaplastic thyroid cancer: molecular pathogenesis and emerging therapies. Endocrine-Related Cancer 16: 17–44.

3. Montero-Conde C, Martin-Campos JM, Lerna E, Gimenez G, Martinez-Guitarte JL, et al. (2008) Molecular profiling related to poor prognosis in thyroid carcinoma. Combining gene expression data and biological information. Oncogene 27: 1554–1561.

4. Salvatore G, Nappi TC, Salerno P, Jiang Y, Garbi C, et al. (2007) A cell proliferation and chromosomal instability signature in anaplastic thyroid carcinoma. Cancer Research 67: 10184–10185.

5. Hebrant A, Dom G, Dewaele M, Andry G, Tresallet C, et al. (2012) mRNA Expression in Papillary and Anaplastic Thyroid Carcinoma: Molecular Anatomy of a Killing Switch. PloS One 7.

6. Li B-Q, Hu L-L, Chen L, Feng K-Y, Cai Y-D, et al. (2012) Prediction of Protein Protocols 4: 44–57.

7. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy of machine learning algorithms. Proceedings of the 14th international joint conference on Artificial intelligence- Volume 2.Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc. pp. 1137–1143.

8. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. Journal of Proteome Research 5: 1888–1897.

9. Peng HC, Long FH, Ding C (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 1226–1236.

10. Friedman JH, Baskett F, Shustek LJ (1975) An Algorithm for Finding Nearest Neighbors. Computers, IEEE Transactions on C-24: 1000–1006.

11. Chou K-C, Shen H-B (2006) Predicting eskyratin protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. Journal of Proteome Research 5: 1080–1087.

12. Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273: 236–247.

13. Li B-Q, Hu L-L, Chen L, Feng K-Y, Cai Y-D, et al. (2012) Prediction of Protein Domain with mRMR Feature Selection and Analysis. PLoS ONE 7: e39308.

14. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. Journal of Proteome Research 5: 1888–1897.

15. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence- Volume 2.Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc. pp. 1137–1143.

16. Dijkstra E (1959) A note on two problems in connection with graphs. Numerische Mathematik 1: 269–271.

17. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4: 44–57.

18. Cronin SJ, Penninger JM (2007) From T-cell activation signals to signaling control of anti-cancer immunity. Immunological Reviews 209: 151–160.

19. Heubner M, Wimberger P, Otterbach F, Kasimir-Bauer S, Siffert W, et al. (2003) Association of the AA genotype of the BCL2 (-938C>A) promoter polymorphism with better survival in ovarian cancer. Int J Biol Markers 24: 223–229.

20. Bachmann HS, Otterbach F, Callies R, Nuckel H, Bartsch M, et al. (2007) The AA genotype of the regulatory BCL2 promoter polymorphism (938C>A) is associated with a favorable outcome in lymph node negative invasive breast cancer patients. Clin Cancer Res 13: 5790–5797.

21. Bachmann HS, Heukamp LC, Schmitz KJ, Hilsenrath C, Kahl P, et al. (2011) Regulatory BCL2 promoter polymorphism (938C>A) is associated with adverse outcome in patients with prostate carcinoma. Int J Cancer 129: 2390–2399.

22. Ross D, Rasi S, Capello D, Gaidano G (2008) Prognostic assessment of BCL2-938C>A polymorphism in chronic lymphocytic leukemia. Blood 111: 466–468.

23. Aksoy M, Giles Y, Kapran Y, Terezcolu T, Tezcan S (2003) Expression of bcl-2 in papillary thyroid carcinomas and its prognostic value. Acta Chir Belg 103: 644–648.

24. Eun YG, Hong IK, Kim SK, Park HK, Kwon S, et al. (2011) A Polymorphism (rs1101018, Thr7Thr) of BCL2 is Associated with Papillary Thyroid Cancer in Korean Population. Clin Exp Otorhinolaryngol 4: 149–154.

25. Noetzel E, Veeck J, Niederacher D, Galm O, Horn F, et al. (2008) Promoter methylation-associated loss of ID4 expression is a marker of tumour recurrence in human breast cancer. BMC Cancer 8: 154.

26. Deluc S, Savonet V, Behrendt J, Dumont JE, Maenhaut C (2002) Study of gene expression in thyrotopin-stimulated thyroid cells by cDNA expression array: ID3 transcription modulating factor as an early response protein and tumor marker in thyroid carcinomas. Exp Cell Res 279: 62–70.

27. Kebebew E, Peng M, Treseler PA, Clark OH, Duh QY, et al. (2004) Id1 gene expression is up-regulated in hyperplastic and neoplastic thyroid tissue and regulates growth and differentiation in thyroid cancer cells. J Clin Endocrinol Metab 89: 6105–6111.

28. Rajolet M, Rich T, Gross-Morand MS, Molina-Gomes D, Viejas-Peiquignon E, et al. (1998) cDNA cloning, tissue distribution and chromosomal localization of the human ID4 gene. DNA Res 5: 309–313.

29. Chen GG, Vlantis AC, Zeng Q, van Hasselt CA (2008) Regulation of cell growth by estrogen signaling and potential targets in thyroid cancer. Curr Cancer Drug Targets 8: 367–377.

30. Fujimoto J, Hirose R, Ichigo S, Sakaguchi H, Tamaya T (1998) DNA polymorphism in B-domain of the estrogen receptor-alpha among Japanese women. Steroids 63: 146–148.

31. Lehrer SP, Schmutzel RK, Rabin JM, Schachter BS (1993) An estrogen receptor genetic polymorphism and a history of spontaneous abortion—correlation in women with estrogen receptor positive breast cancer but not in women with estrogen receptor negative breast cancer or in women without breast cancer. Breast Cancer Res Treat 26: 173–180.

32. Massari F, Becherini L, Gennari L, Facchini V, Genazzani AR, et al. (2001) Association of RAS point mutations and PAX8-PPAR gamma rearrangement in post-Chernobyl thyroid cancer and their association with iodine-131 radiation dose and other characteristics. Cancer 119: 1792–1799.

33. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.

34. Eberhardt NL, Grebe SK, McIver B, Reddi HV (2010) The role of the PAX8/PPARgamma chromosomal rearrangements in post-Chernobyl thyroid cancer and their association with iodine-131 radiation dose and other characteristics. Cancer 119: 1792–1799.

35. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.

36. Eberhardt NL, Grebe SK, McIver B, Reddi HV (2010) The role of the PAX8/PPARgamma chromosomal rearrangements in post-Chernobyl thyroid cancer and their association with iodine-131 radiation dose and other characteristics. Cancer 119: 1792–1799.

37. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.

38. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.

39. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.

40. McViber B, Grebe SK, Eberhardt NL (2004) The PAX8/PPAR gamma fusion oncogene as a potential therapeutic target in follicular thyroid carcinoma. Curr Drug Targets Immune Endocr Metabol Disord 4: 221–234.