

RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo deste relatório de pós-doutorado será disponibilizado somente a partir de 22/12/2025.

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
INSTITUTO DE BIOCÊNCIAS DE BOTUCATU

Dr. André Luiz Molan

**Construção de um Modelo Preditivo Baseado em Assinaturas
Transcricionais para a Caracterização do Risco em Pacientes com Câncer
de Tireoide**

Relatório Científico Circunstanciado exigido para a
conclusão de Pós-Doutorado exigido pelo Programa de
Pós-Doutorado da Unesp.

Grande Área: Ciências Biológicas

Área: Genética

Subárea: Genética Humana e Médica

Supervisor: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Dezembro de 2023.

Resumo

A tireoide é uma glândula endócrina que desempenha diferentes papéis relacionados tanto ao desenvolvimento de órgãos quanto ao controle homeostático de mecanismos fisiológicos relacionados ao crescimento e consumo de energia em todos os vertebrados. A incidência do câncer de tireoide, de forma global, vem aumentando gradativamente. Nos últimos 30 anos, países como os Estados Unidos tiveram o número de casos quase triplicado. O câncer de tireoide é a malignidade mais comum em órgãos endócrinos. A maioria de seus tumores são derivados de células epiteliais foliculares, enquanto 3-5% tem origem parafolicular ou células C (carcinoma medular). Normalmente, ocorrem em nódulos tireoidianos e são detectados por exames de toque ou imagem. Os nódulos, tipicamente, são unilaterais e indolores. Devido à grande maioria dos nódulos serem benignos, o desafio clínico consiste em identificar de forma rápida e precisa os nódulos que realmente abrigam o câncer. A amostragem de nódulos através da biópsia com agulha fina de aspiração (FNA) guiada por ultrassom, seguido de uma análise citológica das células coletadas, são os métodos mais utilizados e que melhores resultados apresentam. Um método alternativo à aspiração citológica é a mini biópsia *trucut*. Ela permite ao patologista ter acesso ao núcleo do tecido tireoidiano. Sua realização consiste no uso de uma agulha *trucut* 18G acompanhado de anestesia local e uso de ultrassom, e normalmente é aplicado quando a técnica com FNA mostra resultados não diagnóstico ou suspeitos de malignidade. Contudo, ambos os métodos são passíveis de falhas. Experimentos utilizando tecnologia de sequenciamento NGS (*Next Generation Sequence*) se tornaram uma grande ferramenta para a melhoria de acurácia nos diagnósticos de câncer de tireoide. Na prática clínica, experimentos NGS tem sido usados, ainda, no diagnóstico de nódulos tireoidianos com citologia indeterminada. Diante deste cenário, com uma grande oferta de dados e grandes avanços tecnológicos, técnicas estatísticas e modelos computacionais se tornaram ferramentas indispensáveis para a análise genômica e transcriptômica tumoral. Com isso, destacamos o aprendizado de máquina, um ramo da inteligência artificial composto por algoritmos capazes de buscar por padrões ou tendências em um dado conjunto de dados e utilizar o conhecimento adquirido para a realização de previsões e busca por novos *insights*.

Abstract

The thyroid is an endocrine gland that plays different roles in the development of organs and the homeostatic control of physiological mechanisms related to growth and energy consumption in all vertebrates. The incidence of thyroid cancer, globally, has been increasing gradually. In the last 30 years, countries like the United States have had the number of cases almost tripled. Thyroid cancer is the most common malignancy of endocrine organs. Most of their tumors are derived from follicular epithelial cells, while 3-5% have parafollicular or C-cell origin (medullary carcinoma). They usually occur in thyroid nodules and are detected by touch or imaging tests. Nodules are typically unilateral and painless. Because the vast majority of nodules are benign, the clinical challenge is to quickly and accurately identify the nodules that actually harbor cancer. Sampling of nodules through ultrasound-guided fine-needle aspiration biopsy (FNA), followed by a cytological analysis of the collected cells, are the most used methods and the ones with the best results. An alternative method to cytological aspiration is the trucut mini biopsy. It allows the pathologist to gain access to the core of the thyroid tissue. Its performance consists of the use of an 18G trucut needle with local anesthesia and use of ultrasound, and is usually applied when the FNA technique shows non-diagnostic results or malignancy suspicious. However, both methods are prone to failure. Experiments using NGS sequencing technology have become a great tool for improving the accuracy of thyroid cancer diagnoses. In clinical practice, NGS experiments have also been used in the diagnosis of thyroid nodules with indeterminate cytology. In this scenario, with a large supply of data and major technological advances, statistical techniques and computational models have become indispensable tools for tumor genomics and transcriptomics analysis. Thus, we highlight machine learning, a branch of artificial intelligence composed of algorithms capable of searching for patterns or trends in a given data set and using the acquired knowledge to make predictions and search for new insights.

Lista de Figuras

1.1	Modelo de progressão	p. 5
4.1	Curva Regressão Logística	p. 10
4.2	<i>Multilayer Perceptron</i>	p. 11
4.3	Árvore de Decisão	p. 11
4.4	Exemplo de Árvore de Decisão	p. 12
4.5	Modelo SVM	p. 12
4.6	Validação Cruzada	p. 13
4.7	Balanceamento amostral	p. 16

Lista de Tabelas

1.1	Tipos de Tumores e Perfis de Mutação	p. 2
1.2	Tipos de Tumores e Perfis de Mutação	p. 3
1.3	Classificação citológica	p. 4
1.4	Estratificação de Risco	p. 5
A.1	Transcritos Assinatura Modelo 1	p. 26
A.2	Transcritos Assinatura Modelo 2	p. 28
B.1	Relação de modelos de aprendizado de máquina obtidos	p. 32
C.1	Enriquecimento Funcional baseado no <i>Gene Ontology</i> Modelo 1	p. 111
C.2	Enriquecimento Funcional baseado no <i>Gene Ontology</i> Modelo 2	p. 121
D.1	Enriquecimento NCG Modelo 1	p. 133
D.2	Enriquecimento DO Modelo 1	p. 134
D.3	Enriquecimento DGN Modelo 1	p. 135

Sumário

Resumo	p. ii
Abstract	p. iii
1 Introdução	p. 1
2 Justificativa	p. 7
3 Objetivos	p. 8
3.1 Objetivos Específicos	p. 8
4 Material e Métodos	p. 9
4.1 Dados Tumoriais	p. 9
4.2 Técnicas de Aprendizado de Máquina	p. 9
4.2.1 Regressão Logística	p. 9
4.2.2 Rede neural com <i>Multilayer Perceptron</i> (MLP)	p. 10
4.2.3 Árvore de Decisão	p. 11
4.2.4 <i>Support Vector Machine</i>	p. 12
4.2.5 Validação Cruzada (<i>Cross Validation</i>)	p. 13
4.3 Seleção de <i>Features</i>	p. 13
4.3.1 Metanálise	p. 13
4.3.2 Método Univariado	p. 14
4.3.3 Expressão Diferencial	p. 14
4.4 Balanceamento Amostral	p. 15
4.5 Construção dos Modelos	p. 16

4.6	Enriquecimento Funcional	p. 17
5	Resultados e Discussões	p. 19
6	Conclusão	p. 21
	Referências Bibliográficas	p. 22
	Apêndice A	p. 25
	Apêndice B	p. 31
	Apêndice C	p. 110
	Apêndice D	p. 132

Referências Bibliográficas

- ALEXA, A.; RAHNENFÜHRER, J. Gene set enrichment analysis with topgo. *Bioconductor Improv*, v. 27, 2009.
- BADR, W. *Having an imbalanced dataset? Here is how you can fix it*. 2019. <https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb>. Acessado em: 20-08-2022.
- BAELDUNG. *Multiclass Classification Using Support Vector Machines*. 2021. <https://www.baeldung.com/cs/svm-multiclass-classification#:~:text=In%20its%20most%20simple%20type,into%20multiple%20binary%20classification%20problems>. Acessado em: 19-09-2022.
- BALI, R. et al. *R: Unleash Machine Learning Techniques*. [S.l.]: Packt Publishing Ltd, 2016.
- BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004.
- BELFIORE, A. et al. Cancer risk in patients with cold thyroid nodules: relevance of iodine intake, sex, age, and multinodularity. *The American journal of medicine*, Elsevier, v. 93, n. 4, p. 363–369, 1992.
- BENVENGA, S. et al. Thyroid gland: anatomy and physiology. *Encyclopedia of Endocrine Diseases*, v. 4, p. 382–390, 2018.
- CARLSON, M. et al. org. hs. eg. db: Genome wide annotation for human. *R package version*, v. 3, n. 2, p. 3, 2019.
- CHA, Y. J.; KOO, J. S. Next-generation sequencing in thyroid cancer. *Journal of translational medicine*, Springer, v. 14, n. 1, p. 1–10, 2016.
- COOPER, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of personality and social psychology*, American Psychological Association, v. 37, n. 1, p. 131, 1979.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.
- IBM. *What is a decision tree?* 2022. <https://www.ibm.com/topics/decision-trees>. Acessado em: 19-08-2022.

- INSTITUTE, E. B. *Differential Expression analysis*. 2022. <https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/differential-gene-expression-analysis/#:~:text=Differential%20expression%20analysis%20means%20taking,expression%20levels%20between%20experimental%20groups>. Acessado em: 05-11-2022.
- LANGFELDER, P.; HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, Springer, v. 9, n. 1, p. 559, 2008.
- MAENHAUT, C. et al. Ontogeny, anatomy, metabolism and physiology of the thyroid. *Endotext [Internet]*, MDText. com, Inc., 2015.
- MEYER, D.; WIEN, F. Support vector machines. *The Interface to libsvm in package e1071*, v. 28, p. 20, 2015.
- MICHALOS, A. C. *Encyclopedia of quality of life and well-being research*. [S.l.]: Springer Netherlands Dordrecht, 2014. 302-308 p.
- MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: IEEE. *2020 11th international conference on information and communication systems (ICICS)*. [S.l.], 2020. p. 243–248.
- NAQA, I. E.; MURPHY, M. J. What is machine learning? In: *machine learning in radiation oncology*. [S.l.]: Springer, 2015. p. 3–11.
- NICK, T. G.; CAMPBELL, K. M. Logistic regression. *Topics in biostatistics*, Springer, p. 273–301, 2007.
- NIKIFOROV, Y. E.; NIKIFOROVA, M. N. Molecular genetics and diagnosis of thyroid cancer. *Nature Reviews Endocrinology*, Nature Publishing Group, v. 7, n. 10, p. 569–580, 2011.
- NIX, P.; NICOLAIDES, A.; COATESWORTH, A. Thyroid cancer review 1: presentation and investigation of thyroid cancer. *International journal of clinical practice*, Wiley Online Library, v. 59, n. 11, p. 1340–1344, 2005.
- NORIEGA, L. Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, Citeseer, 2005.
- PEIXOTO, F. *A simple overview of Multilayer Perceptron (MLP)*. [S.l.]: Analytics Vidhya, 2020. <https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/#>. Acessado em: 19-08-2022.
- ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, Oxford University Press, v. 26, n. 1, p. 139–140, 2010.
- SHAH, J. P. Thyroid carcinoma: epidemiology, histology, and diagnosis. *Clinical advances in hematology & oncology: H&O*, NIH Public Access, v. 13, n. 4 Suppl 4, p. 3, 2015.

SHI, M. et al. Upregulated mitosis-associated genes cenpe, cenpf, and dlga5 predict poor prognosis and chemotherapy resistance of acute myeloid leukemia. *Cancer Biomarkers*, IOS Press, v. 35, n. 1, p. 11–25, 2022.

SMYTH, G. K. Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. [S.l.]: Springer, 2005. p. 397–420.

SPICEWORKS. *What is Logistic Regression? Equation, Assumptions, Types and Best Practices*. 2022. <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>. Acessado em: 19-09-2022.

SPRENT, P. Fisher exact test. In: *International encyclopedia of statistical science*. [S.l.]: Springer, 2011. p. 524–525.

STOUFFER, S. A. et al. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. Princeton Univ. Press, 1949.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 43, p. 15545–15550, 2005.

TOMCZAK, K.; CZERWIŃSKA, P.; WIZNEROWICZ, M. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, Termedia, v. 2015, n. 1, p. 68–77, 2015.

WU, Y. et al. Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. *Frontiers in endocrinology*, Frontiers Media SA, v. 11, p. 577537, 2020.

YANG, C. Q. et al. Creating prognostic systems for well-differentiated thyroid cancer using machine learning. *Frontiers in endocrinology*, Frontiers Media SA, v. 10, p. 288, 2019.

YU, G. et al. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, Oxford University Press, v. 31, n. 4, p. 608–609, 2015.