UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escuela Técnica Superior de Ingeniería Informática
Universidad Politécnica de Valencia

Diseño e implementación de herramientas de análisis de genoma basadas en la teoría de la información

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Cristina Rodríguez Fernández

Tutor: Jose María Sempere Luna

Curso 2024-2025

Resumen

La retinosis pigmentaria (RP) es una de las distrofias hereditarias de la retina (DHR) más comunes, caracterizada por una degeneración progresiva de los fotorreceptores que conduce a una pérdida de visión irreversible. A pesar de los avances en la secuenciación masiva, el diagnóstico genético sigue siendo un desafío debido a la gran heterogeneidad genética de la enfermedad. Aunque se han identificado más de 250 genes asociados a DHR, un porcentaje significativo de pacientes sigue sin un diagnóstico preciso tras el análisis de paneles de genes o incluso del exoma completo (WES).

En este Trabajo de Fin de Grado, que amplía el Trabajo de Fin de Máster de Andrea Vañó Ribelles y Luis, se propone el uso de herramientas de teoría de la información y modelos de cadenas de Markov para el análisis del genoma humano con el objetivo de predecir posibles mutaciones asociadas a la RP. Dado que la RP es una enfermedad rara y el número de individuos afectados es limitado, las técnicas de machine learning no resultan del todo efectivas debido a la escasez de datos para entrenar modelos robustos. En su lugar, este estudio se centra en el análisis de regiones genómicas con alta entropía y elevada densidad de mutaciones, ya que estas zonas pueden ser clave para identificar variantes con mayor probabilidad de estar relacionadas con la patología.

Para ello, se emplearán datos genómicos disponibles en el National Center for Biotechnology Information (NCBI) y archivos VCF proporcionados por el Grupo de Investigación del IIS La Fe de Biomedicina Molecular, Celular y Genómica. Con esta aproximación, se busca mejorar la comprensión de los mecanismos subyacentes de la enfermedad y contribuir al desarrollo de herramientas más eficaces para el diagnóstico genético de la RP.

Palabras clave: retinosi pigmentària, dades genòmiques, teoria de la informació, cadenes de Markov, predicció de mutacions, machine learning, entropia

Resum

La retinosi pigmentària (RP) és una de les distròfies hereditàries de la retina (DHR) més comunes, caracteritzada per una degeneració progressiva dels fotoreceptors que condueix a una pèrdua de visió irreversible. Malgrat els avanços en la seqüenciació massiva, el diagnòstic genètic continua sent un repte a causa de la gran heterogeneïtat genètica de la malaltia. Encara que s'han identificat més de 250 gens associats a DHR, un percentatge significatiu de pacients continua sense un diagnòstic precís després de l'anàlisi de panells de gens o fins i tot de l'exoma complet (WES).

En aquest Treball de Fi de Grau, que amplia el Treball de Fi de Màster d'Andrea Vañó Ribelles i Luis, es proposa l'ús d'eines de teoria de la informació i models de cadenes de Markov per a l'anàlisi del genoma humà amb l'objectiu de predir possibles mutacions associades a la RP. Com que la RP és una malaltia rara i el nombre d'individus afectats és limitat, les tècniques de machine learning no resulten del tot efectives a causa de l'escassetat de dades per a entrenar models robustos. En el seu lloc, aquest estudi se centra en l'anàlisi de regions genòmiques amb alta entropia i elevada densitat de mutacions, ja que aquestes zones poden ser clau per a identificar variants amb més probabilitat d'estar relacionades amb la patologia.

Per a això, s'empraran dades genòmiques disponibles en el National Center for Biotechnology Information (NCBI) i arxius VCF proporcionats pel Grup d'Investigació de l'IIS La Fe de Biomedicina Molecular, Cel·lular i Genòmica. Amb aquesta aproximació,

es busca millorar la comprensió dels mecanismes subjacents de la malaltia i contribuir al desenvolupament d'eines més eficaces per al diagnòstic genètic de la RP.

Paraules clau: retinosis pigmentaria, datos genómicos, teoría de la información, cadenas de Markov, predicción de mutaciones, machine learning, entropía

Abstract

Retinitis pigmentosa (RP) is one of the most common hereditary retinal dystrophies (HRD), characterized by a progressive degeneration of photoreceptors leading to irreversible vision loss. Despite advances in massive sequencing, genetic diagnosis remains challenging due to the high genetic heterogeneity of the disease. Although more than 250 genes have been associated with HRD, a significant percentage of patients remain undiagnosed even after gene panel analysis or whole-exome sequencing (WES).

This Bachelor's Thesis, which builds upon the Master's Thesis by Andrea Vañó Ribelles and Luis, proposes the use of information theory tools and Markov chain models for human genome analysis to predict potential RP-associated mutations. Since RP is a rare disease and the number of affected individuals is limited, machine learning techniques are not entirely effective due to the lack of sufficient data to train robust models. Instead, this study focuses on analyzing genomic regions with high entropy and a high density of mutations, as these areas may be key to identifying variants with a greater likelihood of being related to the disease.

To achieve this, genomic data from the National Center for Biotechnology Information (NCBI) and VCF files provided by the Research Group at IIS La Fe for Molecular, Cellular, and Genomic Biomedicine will be used. This approach aims to improve the understanding of the underlying mechanisms of the disease and contribute to the development of more effective tools for the genetic diagnosis of RP.

Key words: retinitis pigmentosa, genomic data, information theory, Markov chains, mutation prediction, machine learning, entropy

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducció

????? ????????????? ????????????? ????????????? ????????????? ?????????????

1.1 Motivació

????? ????????????? ????????????? ????????????? ????????????? ?????????????

1.2 Objectius

????? ????????????? ????????????? ????????????? ????????????? ?????????????

1.3 Estructura de la memòria

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 2
??? ???? ??????

????? ????????????? ????????????? ????????????? ????????????? ?????????????

2.1 ?? ???? ???? ? ?? ??

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 3
??? ???? ??????

????? ????????????? ????????????? ????????????? ????????????? ?????????????

3.1 ?? ???? ???? ? ?? ??

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 4
Conclusions

????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

[1] Jennifer S. Light. When computers were women. *Technology and Culture*, 40:3:455–483, juliol, 1999.

[2] Georges Ifrah. *Historia universal de las cifras*. Espasa Calpe, S.A., Madrid, sisena edició, 2008.

[3] Comunicat de premsa del Departament de la Guerra, emés el 16 de febrer de 1946. Consultat a http://americanhistory.si.edu/comphist/pr1.pdf.

APÉNDICE A
Configuració del sistema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.1 Fase d'inicialització

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.2 Identificació de dispositius

????? ????????????? ????????????? ????????????? ????????????? ?????????????

APÉNDICE B
??? ????????????? ????

????? ????????????? ????????????? ????????????? ????????????? ?????????????