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 retina (DHR) más comunes, esta se caracteriza por una degeneración progresiva de los fotorreceptores que conduce a una pérdida de visión irreversible. A pesar de los avances en la secuenciación del genoma humano, el origen de esta condición sigue siendo incierto debido a la gran heterogeneidad genética de la enfermedad y la cantidad de genes asociados a DHR.

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 Fuentes 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 lo suficientemente efectivas debido a la escasez de datos para entrenar modelos robustos. En su lugar, este enfoque 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. El propósito de este estudio es profundizar en el conocimiento de las características del genoma asociadas a la enfermedad y aportar nuevas estrategias que ayuden a optimizar el diagnóstico genético de la retinosis pigmentaria.

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 retina (DHR) més comunes. Es caracteritza per una degeneració progressiva dels fotoreceptors que provoca una pèrdua de visió irreversible. Malgrat els avanços en la seqüenciació del genoma humà, l'origen d'aquesta malaltia continua sent incert a causa de la gran heterogeneïtat genètica de la patologia i la quantitat de gens associats a les DHR.

En aquest Treball de Fi de Grau, que amplia el Treball de Fi de Màster d'Andrea Vañó Ribelles i Luis, es proposa la utilització d'eines de la Teoria de la Informació i models de Fonts 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 de persones afectades és reduït, les tècniques de machine learning no resulten prou efectives a causa de l'escassetat de dades per a entrenar models robustos. En lloc d'això, aquest enfocament 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 una major probabilitat d'estar relacionades amb la patologia.

Per a dur a terme aquest estudi, s'utilitzaran dades genòmiques disponibles al 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. L'objectiu d'aquest treball és aprofundir en el coneixement de les característiques del genoma associades a la malaltia i aportar noves estratègies que ajuden a optimitzar el diagnòstic genètic de la retinosi pigmentària.

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). It is characterized by the progressive degeneration of photoreceptors, leading to irreversible vision loss. Despite advances in human genome sequencing, the origin of this condition remains uncertain due to the high genetic heterogeneity of the disease and the large number of genes associated with HRD.

In this Bachelor's Thesis, which expands upon the Master's Thesis by Andrea Vañó Ribelles and Luis, the use of Information Theory tools and Markov Source models is proposed for the analysis of the human genome, with the aim of predicting possible mutations associated with RP. Since RP is a rare disease and the number of affected individuals is limited, machine learning techniques are not sufficiently effective due to the lack of data needed to train robust models. Instead, this approach focuses on the analysis of genomic regions with high entropy and a high density of mutations, as these areas may be key to identifying variants with a higher probability of being related to the pathology.

To achieve this, genomic data available from the National Center for Biotechnology Information (NCBI) and VCF files provided by the Research Group of the IIS La Fe in Molecular, Cellular and Genomic Biomedicine will be used. The purpose of this study is to deepen the understanding of the genomic features associated with the disease and to provide new strategies that help optimize the genetic diagnosis of retinitis pigmentosa.

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
Introducción

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

1.1 Motivación

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

1.2 Objetivos

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

1.3 Estructura de la memoria

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

CAPÍTULO 2
??? ???? ??????

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

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

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

CAPÍTULO 3
??? ???? ??????

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

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

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

CAPÍTULO 4
Conclusiones

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

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

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