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), una de las distrofias hereditarias de retina (DHR) más frecuente, es conocida por una degradación progresiva en los fotorreceptores que termina por causar pérdida visual irreversible. Aún con los recientes avances en la secuenciación del genoma humano, el origen de esta condición sigue siendo un desafío debido a la amplia diversidad genética involucrada y la cantidad de genes relacionados con las DHR.

Este Trabajo de Fin de Grado, continuación del Trabajo de Fin de Máster de Andrea Vañó Ribelles y el Trabajo de Fin de Grado de Luis Alberto Martínez Bravo, explora el uso de herramientas de Teoría de la Información y modelos de Fuentes de Markov para analizar el genoma humano, con el objetivo de predecir posibles mutaciones asociadas a la RP. Puesto que la RP es una enfermedad rara y el número de personas afectadas es muy pequeño, las técnicas de Machine Learning no resultan lo suficientemente efectivas debido a la falta de datos para desarrollar modelos robustos.

En vez de ello, este enfoque se centra en el análisis de regiones genómicas con elevada entropía y densidad de mutaciones, ya que estas áreas serían clave para detectar variantes con mayor probabilidad de estar vinculadas a la enfermedad.

Para ello, se han empleado los datos genómicos del 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 del estudio es profundizar en las características del genoma asociadas a la enfermedad y aportar nuevas maneras de optimizar el diagnóstico genético de la retinosis pigmentaria.

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

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

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

El objetivo principal de este Trabajo de Fin de Grado es proporcionar herramientas informáticas para el análisis genómico, en concreto aplicado a enfermedades raras como la Retinosis Pigmentaria (RP). Para ello contamos con archivos que contienen información sobre mutaciones y secuencias completas del genoma, y se implementan modelos basados en Teorías de la Información y Fuentes de Markov, con el fin de identificar regiones genómicas relevantes y predecir mutaciones potencialmente patogénicas.

1.2.1. Procesamiento de archivos VCF y secuencias genómicas

Hacer uso de herramientas y librerías de uso biomédico que permitan la lectura, escritura y análisis de archivos genéticos, y de esta manera, realizar un preproceso y de la información genética y aplicar mutaciones sobre la secuencia de referencia de forma precisa y eficiente.

Preguntas de Investigación:

- ¿Qué formato presentan los archivos VCF y cómo se puede garantizar la compatibilidad con los archivos FASTA con la secuencia genómica?
- ¿Cómo se pueden integrar las trasformaciones necesarias en la secuencia de referencia de manera correcta y sin conflictos entre ellas?

1.2.2. Aplicación de Teorías de la Información y Fuentes de Markov al análisis de secuencias genómicas

Utilizar modelos de Teorías de la Información y Fuentes de Markov para predecir el impacto de las mutaciones en la secuencia genómica y así detectar zonas con valores anómalos de entropía, o alta densidad de mutaciones, para su posterior estudio.

Preguntas de Investigación:

- ¿Qué diferencias se observan en los valores de entropía, tanto la simple como la realizada a partir de aplicar Fuentes de Markov, entre la secuencia original y la mutada?
- ¿Hay alguna correlación entre las zonas con alta densidad de mutaciones y los valores de entropía?
- ¿Cómo es la estructura que presenta el autómata de transición generado a partir de Fuentes de Markov de orden k?
- ¿Cuáles son los valores óptimos de los parámetros para obtener resultados interesantes?

1.2.3. Desarrollo de una interfaz de usuario para la visualización de resultados

Crear una aplicación para facilitar el uso del programa, sin necesidad de tener conocimientos avanzados en informática. Esto es especialmente importante en este contexto, ya que se trata de una herramienta enfocada a un usuario que no es experto en informática. Además, esta interfaz debe permitir la interacción del usuario con la aplicación, modificando distintos parámetros para obtener diferentes resultados en función de lo que se busque, y visualizar gráficamente y de manera unificada los resultados de este análisis.

Preguntas de Investigación:

- ¿Qué diseño debe tener la interfaz para que no sea demasiado compleja, pero permita interactuar ampliamente al usuario?
- ¿Cómo presentar los resultados de manera ordenada, clara y concisa, para hacer accesible la información más complicada?

1.2.4. Contribución a las aplicaciones de la IA en el campo de la investigación genética

Proponer un nuevo enfoque de análisis genómico basado en Teorías de la Información, el cual no se había estudiado aún, y ofrecer una herramienta efectiva para el estudio de otras enfermedades raras con componente genético.

Preguntas de Investigación:

- ¿Debe poder generalizarse el proyecto para otras patologías?
- ¿Qué técnicas se podrían utilizar sobre los resultados obtenidos, como machine learning, para clasificar o encontrar patrones de mutaciones?

1.3 Contexto

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

1.4 Estructura de la memoria

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

CAPÍTULO 2
Estado del arte

El campo del análisis genómico de la retinosis pigmentaria (RP) está avanzando rápidamente, con desarrollos en la identificación genética, herramientas de predicción y terapias emergentes. Sin embargo, en un contexto de disponibilidad limitada de especialistas en retina, pruebas y asesoramiento genético, sigue existiendo una gran necesidad de métodos diagnósticos precisos y accesibles. Esta situación ha motivado la búsqueda de nuevas técnicas para mejorar la detección de esta enfermedad.

2.1 Métodos tradicionales

Los métodos tradicionales basados en sistemas de codificación médica como el ICD (International Classification of Diseases) y el SNOMED (Systematized Nomenclature of Medicine) han sido ampliamente utilizados en la medicina para clasificar enfermedades y registrar diagnósticos. Sin embargo, en el caso de enfermedades genéticas raras como la retinosis pigmentaria (RP), estos sistemas de codificación presentan importantes limitaciones[3]. Esto es debido a que es una patología altamente heterogénea a nivel genético y clínico: existen numerosas mutaciones responsables y los síntomas pueden evolucionar de formas muy diversas entre pacientes[4].

2.2 Inteligencia Artificial

En la actualidad se están utilizando métodos basados en inteligencia artificial (IA) para la detección, diagnóstico y pronóstico de numerosas enfermedades en distintas áreas de la medicina, desde oncología hasta neurología o cardiología. Esto es debido a la capacidad de la IA para trabajar con grandes volúmenes de datos e identificar en estos patrones complejos para generar predicciones con una alta precisión[5]. No obstante, su desarrollo para las distrofias hereditarias de retina (DHR), específicamente la retinosis pigmentaria (RP) todavía está en una fase temprana.

2.2.1. Aplicaciones de Machine Learning

El aprendizaje profundo (deep learning) es una subcategoría de la inteligencia artificial que ha ganado mucha atención en los últimos años, especialmente porque el aprendizaje profundo es muy eficaz en el reconocimiento de patrones y el análisis de imágenes[6].

En un estudio reciente, se utilizaron tres modelos preentrenados; Inception-v3, ResNet-50 y VGG-19, para clasificar imágenes retinianas asociadas a diferentes genes relaciona-

dos con la retinosis pigmentaria. Tras un preprocesamiento exhaustivo que incluyó técnicas de class balancing y boosting para corregir la variabilidad genética, los modelos obtuvieron precisiones superiores al 80 % en los datos de training[5]. Sin embargo, al aplicar estos modelos a datos de testing, las tasas de precisión cayeron a un 54 %, 56 % y 54 % respectivamente, resultados claramente insuficientes para garantizar un diagnóstico clínico fiable debido a la elevada tasa de error.

Los estudios de IA disponibles, como el mencionado anteriormente, que buscan la detección, clasificación y predicción de DHR, siguen siendo en su mayoría retrospectivos e incluyen un número relativamente limitado de pacientes debido a su escasez[7]. Esto pone de manifiesto la necesidad de continuar optimizando las metodologías empleadas para alcanzar niveles de exactitud que resulten aceptables en el ámbito médico.

2.2.2. Aplicaciones de Teorías de la Información y Fuentes de Markov

Aunque actualmente no existen estudios específicos que utilicen directamente Teorías de la Información y Fuentes de Markov al análisis genómico de la RP, estos enfoques se han consolidado como potenciales herramientas.

Sabemos que las Teorías de la Información son útiles para medir la cantidad de información, la incertidumbre y el contenido informativo en sistemas complejos, como es el caso del genoma humano. En genómica, conceptos como la entropía permiten detectar mutaciones en el genoma humano, ya que son capaces de identificar cambios en la información genética, comparar con referencias conocidas o analizar patrones[8].

Por su parte, los modelos de Fuentes de Markov (de orden k) permiten capturar patrones de dependencia entre nucleótidos y por lo tanto identificar regiones del genoma con propiedades estadísticas anómalas[9], indicativas de posibles mutaciones patológicas.

2.3 Otras terapias emergentes

Además de los enfoques basados en el análisis genético, en los últimos años también se está abordando el problema desde la estrategia de la terapia genética, dirigida al gen RPGR que es uno de los genes del cromosoma X que comúnmente está asociado a la RP (se tiene constancia de que es el causante de entre el 70 % y el 90 % de los casos ligados al cromosoma X)[10]. Estas terapias consisten en la inyección subretinal de un virus modificado que transporta copias funcionales del gen RPGR, con el objetivo de restaurar la función perdida en las células de la retina afectadas[11].

Sin embargo, para que estas terapias sean verdaderamente efectivas, es esencial identificar de manera precisa las mutaciones responsables. Mediante el uso de herramientas basadas Teorías de la Información es posible identificar de forma eficiente estas variantes, y así seleccionar a los candidatos adecuados para las terapias génicas y optimizar los ensayos clínicos.

CAPÍTULO 3
Fundamentos teóricos

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

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

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

CAPÍTULO 4
Metodología

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

4.1 ?? ???? ???? ? ?? ??

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

CAPÍTULO 5
Desarrollo

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

5.1 ?? ???? ???? ? ?? ??

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

CAPÍTULO 6
Resultados

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

6.1 ?? ???? ???? ? ?? ??

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

CAPÍTULO 7
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] Verana Health. Why Real-World Data is Key to Providing Insights Into Retinitis Pigmentosa, emés el 20 de novembre de 2024. Disponible en https://veranahealth.com/why-real-world-data-is-key-to-providing-insights-into-retinitis-pigmentosa/.

[4] Hartong, D. T., Berson, E. L., & Dryja, T. P. Retinitis pigmentosa. *The Lancet*, vol. 368, núm. 9549, pp. 1795–1809, 2006. Disponible en https://doi.org/10.1016/s0140-6736(06)69740-7.

[5] Ferreira, H., Marta, A., Machado, J., Couto, I., Marques, J. P., Beirão, J. M., & Cunha, A. Retinitis Pigmentosa Classification with Deep Learning and Integrated Gradients Analysis. *Applied Sciences*, vol. 15, núm. 4, art. 2181, 2025. Disponible en https://doi.org/10.3390/app15042181.

[6] Stevenson, S. How deep learning may play a role in predicting retinitis pigmentosa visual impairment, emés el 14 de març de 2023. Disponible en https://www.ophthalmologytimes.com/view/how-deep-learning-may-play-a-role-in-predicting-retinitis-pigmentosa-visual-impairment

[7] Issa, Mohamad et al. Applications of artificial intelligence to inherited retinal diseases: A systematic review. *Survey of Ophthalmology*, vol. 70, núm. 2, pp. 255–264.

[8] Societat Espanyola de Ciències de la Computació i Llenguatges (SECal). Aplicaciones de la Teoría de la Información en Genómica, sense data. Disponible en https://www.secal.es.

[9] N.D. Un algoritmo encuentra indicios de enfermedades en la parte del ADN que teóricamente no servía para nada, emés el 21 d'abril de 2025. Disponible en https://media.lavozdegalicia.es/noticia/sociedad/2025/04/21/algoritmo-encuentra-indicios-enfermedades-parte-adnconsidera-inutil/00031745236367256507381.htm.

[10] Wang, Y., Juroch, K., Chen, Y., Ying, G., & Birch, D. G. Deep Learning–Facilitated Study of the Rate of Change in Photoreceptor Outer Segment Metrics in RPGR-Related X-Linked Retinitis Pigmentosa. *Investigative Ophthalmology & Visual Science*, vol. 64, núm. 14, art. 31, 2023. Disponible en https://doi.org/10.1167/iovs.64.14.31.

[11] ZonaIT. Avances en la terapia génica para la retinosis pigmentaria causada por mutaciones en el gen RPGR, sense data. Disponible en https://biotech-spain.com/es/articles/avances-en-la-terapia-g-nica-para-la-retinosis-pigmentaria-causada-por-mutaciones-en-

APÉNDICE A

Configuració del sistema

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

A.1 Fase d'inicialització

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

A.2 Identificació de dispositius

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

APÉNDICE B
??? ????????????? ????

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