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 Tnformación

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

Apéndices

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

La retinosis o retinitis pigmentaria (RP), una de las distrofias hereditarias de retina DHR más frecuente, es un grupo de enfermedades conocidas por una degradación progresiva en los fotorreceptores (células de la retina responsables de captar la luz y, por tanto, permitir la visión) que termina por causar pérdida visual irreversible. Esta patología suele manifestarse en un principio con ceguera nocturna, seguida por una reducción del campo visual periférico y, en fases avanzadas, puede llegar a la pérdida de la visión central[5]. La RP presenta una gran heterogeneidad genética, lo que significa que puede estar causada por mutaciones en más de 100 genes diferentes, heredados bajo distintos patrones[6]. 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.

1.1 Motivación

1.1.1. Motivación personal

Desde que comencé mi formación en Ingeniería Informática, siempre he tenido la convicción de que esta no debe limitarse a aspectos técnicos, algoritmos o eficiencia computacional, sino que debe orientarse también hacia herramientas que tengan un impacto positivo en la vida de las personas. Por eso, me interesan especialmente las aplicaciones

la Inteligencia Artificial para mejorar la sociedad, hacer el mundo más justo e inclusivo y resolver problemas reales.

Siguiendo estos objetivos, y gracias al proyecto conjunto del Instituto VRAIN con el Grupo de Investigación del IIS La Fe de Biomedicina Molecular, Celular y Genómica, he tenido la oportunidad de formarme en diversas áreas de conocimiento y aplicar mis conocimientos técnicos a una dimensión ética y social. Por esta razón, considero que mejorar la calidad de vida de pacientes mediante herramientas que ayuden a investigadores y personal sanitario en el diagnóstico es una de las principales motivaciones que me impulsa en la realización de este proyecto.

1.1.2. Motivación profesional

Las distrofias hereditarias de retina (DHR) constituyen un conjunto de enfermedades genéticas que afectan a la estructura y función de la retina, llevando progresivamente a la pérdida de visión, en muchos casos hasta alcanzar la ceguera legal. A pesar de ser consideradas enfermedades raras, su impacto es profundo y duradero, no solo desde el punto de vista funcional, sino también psicológico, afectando significativamente la calidad de vida de quienes las padecen[7].

Actualmente, no existe ningún tratamiento curativo, aunque se están explorando alternativas terapéuticas como las que se describen en la sección 2 de esta memoria, pero aún se encuentran en fase experimental. En este contexto, el diagnóstico genético tiene un papel fundamental en la orientación médica y la inclusión del paciente en ensayos clínicos[8].

Uno de los principales desafíos en el estudio genético de las distrofias hereditarias de retina (DHR) radica en la gestión y análisis de la enorme cantidad de variantes obtenidas a partir de técnicas como la secuenciación del exoma completo (WES). Aunque se aplican filtros por frecuencia poblacional o predictores del impacto funcional, el volumen de información sigue siendo elevado, lo que dificulta alcanzar un diagnóstico genético preciso[9].

En este trabajo se propone un enfoque complementario, basado en el uso de modelos computacionales que aplican principios de la teoría de la información y fuentes de Markov para analizar la secuencia genómica antes y después de la introducción de mutaciones. A través de métricas como la entropía y la densidad mutacional, se busca detectar alteraciones estructurales o funcionales que puedan asociarse con variantes patológicas. Esta estrategia no solo contribuye al desarrollo de nuevas herramientas de apoyo al diagnóstico, sino que también tiene una dimensión humana relevante: mejorar el acceso a un diagnóstico certero, lo que representa un paso fundamental para que los pacientes y sus familias encuentren respuestas, orientación clínica y esperanza de acceso a futuras terapias.

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 Teoría 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 Teoría de la Información y Fuentes de Markov al análisis de secuencias genómicas

Utilizar modelos de Teoría 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 la Teoría 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 Estructura de la memoria

La memoria de este TFG está estructurada en 7 secciones, cada una de estas a su vez divididas en subsecciones.

La primera sección se introduce el proyecto, se exponen el contexto y la motivación, y se justifican los objetivos. En la sección 2, se realiza una investigación del conocimiento y tecnologías existentes en la actualidad sobre el tema que se aborda en el proyecto, así como las limitaciones que aún existen, lo que justifica la necesidad de este proyecto. En la sección 3, se realiza una exposición de los conceptos clave necesarios para entender este trabajo, genética y mutaciones, los archivos genómicos utilizados y los principios de la Teoría de la Información. Después, en la sección 4, se describe la metodología seguida y el flujo de información y herramientas usadas a lo largo del proyecto. En la sección 5, se explica detalladamente el funcionamiento del programa. En la sección 6 se exponen y analizan los resultados obtenidos mediante la realización de pruebas cambiando los diferentes parámetros. Por último, en la sección 7, se finaliza el trabajo, explicando las conclusiones obtenidas, seguida por un listado de las referencias bibliográficas utilizadas.

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[10]. 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[11].

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[12]. 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[13].

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[12]. 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[14]. 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 Teoría de la Información y Fuentes de Markov

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

Sabemos que la Teoría de la Información es útil 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[15].

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[16], 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, un tipo de terapia curativa que busca tratar la causa subyacente de una enfermedad genética corrigiendo o reemplazando directamente el gen defectuoso. Esta va 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)[17]. 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[18].

Sin embargo, para que estas terapias sean verdaderamente efectivas, es esencial identificar de manera precisa las mutaciones responsables. Mediante el uso de herramientas basadas en Teoría 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 Conceptos básicos de genética y mutaciones

En el núcleo de cada célula humana se encuentra el ADN, que contiene toda la información genética necesaria para el funcionamiento y desarrollo del organismo. Sin embargo, el ADN no actúa directamente para realizar funciones celulares. En su lugar, esa información debe ser primero transcrita en moléculas de ARN, las cuales desempeñan roles clave en la expresión génica. Este proceso, conocido como transcripción, da lugar a diferentes tipos de ARN, que luego participan en distintos mecanismos celulares.

Existen dos tipos principales de transcritos que resultan de este proceso: el ARN mensajero (ARNm), que lleva la información del ADN a los ribosomas para la síntesis de proteínas, y el ARN no codificante (ARNnc), que, aunque no se traduce en proteínas, cumple funciones esenciales como la regulación génica, la modificación del ARN, la estructura de los ribosomas y otras tareas fundamentales para el control celular.

El ARN mensajero (ARNm) actúa como intermediario entre el ADN y las proteínas de la siguiente manera: mediante el proceso de traducción, los ribosomas convierten cada molécula de ARNm en una proteína, es decir, en una secuencia de aminoácidos. Cada proteína adopta una estructura tridimensional compleja y funciones específicas. Gracias a la interacción entre el ADN, ARN y proteínas, el genoma es capaz de regular qué células deben crecer, morir y cómo se estructuran, entre otras funciones.

Por esta razón, cada individuo presenta un genoma ligeramente distinto al resto. Estas diferencias son conocidas como variaciones o mutaciones genómicas, y pueden suponer variaciones de un único nucleótido o múltiples nucleótidos. Estos cambios influyen en el fenotipo del individuo y también pueden ser responsables de ciertas enfermedades.

- SNV: sustitución de una sola base del ADN por otra. Es el tipo más común de variación genética en el genoma humano.

    - Sinónima: La sustitución no cambia el aminoácido codificado por el codón, debido a la degeneración del código genético. No suele tener efecto funcional, aunque puede afectar la eficiencia de traducción o el splicing.
    - No sinónima: Cambia el aminoácido codificado. Esta a su vez se subdivide en Missense (se sustituye un aminoácido por otro, por lo que puede alterar la función o estructura de la proteína) y Nonsense (introduce un codón de parada prematuro, truncando la proteína, y esta normalmente acaba inactiva o no funcional).

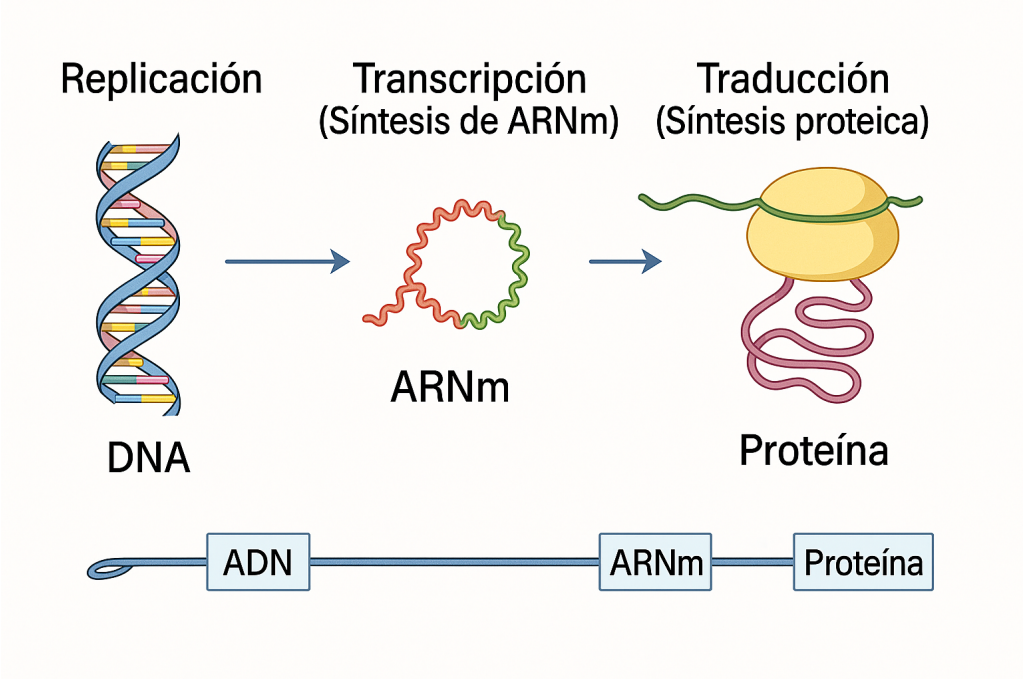

Figura 3.1: Procesos genéticos de la síntesis de proteínas.

- Indels (Inserciones y deleciones): Variaciones que implican la inserción o eliminación de un pequeño número de nucleótidos. Son el segundo tipo más común de variación.
  - In-frame indel: El número de nucleótidos es múltiplo de 3. Esto afecta uno o más codones completos, pero no altera el marco de lectura. Puede tener o no efecto funcional.
  - Frameshift indel: El número de nucleótidos no es múltiplo de 3, por lo que altera el marco de lectura de todo el gen a partir del punto del indel, generando proteínas aberrantes y normalmente no funcionales.
- Copy Number Variation (CNV): Amplificaciones o deleciones de segmentos largos del ADN. Pueden incluir genes completos o regiones reguladoras. Pueden estar relacionadas con la variabilidad fenotípica, trastornos del desarrollo y enfermedades neuropsiquiátricas[19].
- Structural Variation (SV): Alteraciones grandes en la estructura del genoma que incluyen inversiones, translocaciones, inserciones o deleciones mayores.
  - Inversiones: Un segmento se invierte dentro del mismo cromosoma.
  - Translocaciones: Fragmentos de ADN se trasladan a otras posiciones o cromosomas.
  - Grandes deleciones/duplicaciones: afectan múltiples genes/regiones.

3.2 Conceptos básicos de ficheros genéticos

3.3 Teoría de la Información

La teoría de la información es una disciplina que estudia la cuantificación, almacenamiento y comunicación de datos, especialmente enfocada en la transmisión de informa-

ción a través de canales. Fue desarrollada por Claude Shannon y Warren Weaver en los años 40. Sin embargo, su aplicación se ha extendido a otros campos como la informática, la lingüística, o la biología[4].

"Una fuente de información se define como un par (S, P) donde S es un alfabeto predefinido y P es una distribución de probabilidad sobre S. Dado que la fuente de información introduce una incertidumbre en la variable aleatoria definida por el alfabeto predefinido, la entropía mide el grado de incertidumbre de dicha variable. También el grado de aleatoriedad en la fuente de información y, en consecuencia, permite estimar las unidades de información necesarias en promedio para codificar todos los valores posibles que puedan darse en la fuente de información."[2]

Para entender mejor esta idea, definiremos unos conceptos clave:

3.3.1. Canal de comunicación

Un canal de comunicación es el medio físico o digital a través del cual se transmite la información.

3.3.2. Compresión de datos

La compresión hace referencia a la reducción del tamaño de los datos sin perder información.

3.3.3. Código

El código es el sistema de símbolos o señales utilizado para representar la información.

3.3.4. Entropía

El concepto de entropía se introdujo en el siglo XIX en la termodinámica como una magnitud física que media el grado de desorden en un sistema. En la etapa de 1940, Shanon trasladó este concepto al ámbito de las comunicaciones, ya que definió la entropía como una medida de incertidumbre o sorpresa asociada a un conjunto de mensajes[20, p. 4]. Esta idea fundó lo que se conoce hoy en día como Teoría de la Información. Dada una variable aleatoria X con distribución de probabilidad P, la entropía se define como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

La entropía se interpreta como el número promedio de bits necesarios para codificar los resultados de una variable aleatoria.

- Entropía máxima: si una variable aleatoria tiene n resultados igualmente probables, es decir, todos tienen la misma probabilidad de ocurrencia, su entropía vale log 2 n.
- Entropía mínima: si hay certeza de ocurrencia de un evento, es decir, tiene probabilidad 1 y por lo tanto el resto 0, la entropía vale 0. [20, p. 7]

En el contexto de este Trabajo de Fin de Grado, esta idea se aplica a secuencias de ADN, donde cada símbolo representa una base nitrogenada (Adenina, Citosina, Timina

y Guanina) y la probabilidad de ocurrencia de cada una se mide mediante la frecuencia relativa en una región determinada del genoma. Por ejemplo, calcular la entropía de un conjunto de k-mers en una secuencia de ADN podría ayudar a identificar regiones altamente conservadas o variables. Estas regiones pueden correlacionarse con funciones biológicas importantes, como promotores, regiones reguladoras, o sitios de mutación frecuentes.

3.3.5. Modelos de Fuentes de Markov

"Una fuente de información con memoria nula se define como un par (S,P) donde la emisión de cada símbolo s_i sólo depende de su probabilidad $p(s_i)$. Una fuente de información con memoria (de orden m) se define como un par (S,P) de forma que la emisión de cada símbolo s_i depende de los m símbolos emitidos con anterioridad, con una probabilidad condicional $p(s_i \mid s_{i1}, s_{i2}, \ldots, s_{im})$."[2]

Las fuentes de información básicas no tienen memoria, es decir, en el cálculo de la probabilidad de una determinada base no se tienen en cuenta las bases anteriores. Sin embargo, en genómica, el ADN no es del todo aleatorio, ciertas bases tienden a seguir a otras con mayor probabilidad, lo que sugiere que un modelo de Fuente de Markov puede ser más realista.

Una Fuente de Markov (o fuente de información con memoria) se define como un sistema donde la probabilidad de que ocurra cada símbolo depende de los símbolos que le preceden. Estas nuevas probabilidades se consiguen gracias a la matriz de probabilidades de transición entre estados (los estados de una fuente de Markov de orden m serán todas las posibles combinaciones de m símbolos, es decir, n^m). Además, dada una fuente de Markov de orden m, se puede establecer su diagrama de estados, que muestra los estados y las probabilidades condicionales entre ellos[20, p. 20].

- Una fuente de Markov diremos que es ergódica si todos los estados son alcanzables (observables) desde cualquier estado.
- Una fuente de Markov es homogénea si las probabilidades de transición entre estados no cambian a lo largo del tiempo.
- Una fuente de Markov está en estado estacionario si la probabilidad de observación de sus estados no cambia a lo largo del tiempo. Las probabilidades se pueden obtener mediante la ecuación $\pi = \pi \times \Pi$, donde π es el vector de probabilidad de los estados y Π es la matriz de probabilidades condicionales (estocástica).

Por lo tanto, en una fuente de Markov de orden m, la probabilidad de ocurrencia de un símbolo, depende de los (m-1) símbolos anteriores. Cuanto mayor sea el orden (m), más información del contexto de incluye, y, en consecuencia, menores valores de entropía. Esto significa que modificar los valores de m, puede revelar patrones o regiones funcionales dentro de la secuencia dada.

Ejemplo

Tomemos la fuente de Markov de orden 2 definido por el siguiente diagrama de transiciones

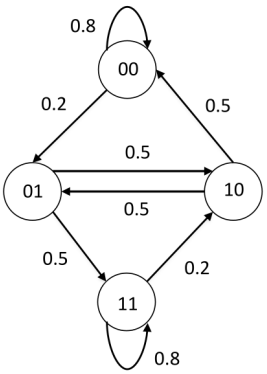

Tabla de probabilidades

$s_j s_k s_i$	$p(s_i \mid s_j s_k)$	$p(s_j s_k)$	$p(s_i, s_j, s_k)$
000	0,8	5/14	4/14
001	0,2	5/14	1/14
010	0,5	2/14	1/14
011	0,5	2/14	1/14
100	0,5	2/14	1/14
101	0,5	2/14	1/14
110	0,2	5/14	1/14
111	0,8	5/14	4/14

$$H(S) = -\sum_{S^{m+1}} p(s_{i_1}, s_{i_2}, \dots, s_{i_m}, s_i) \cdot \log p(s_i \mid s_{i_1}, s_{i_2}, \dots, s_{i_m})$$

$$H(S) = -2 \cdot 4/14 \cdot \log 0{,}8 \; -2 \cdot \; 1/14 \cdot \log 0{,}2 \; -4 \cdot 1/14 \cdot \log 0{,}5 = 0{,}81$$

Figura 3.2: Ejemplo de cálculo de entropía a partir de una Fuente de Markov.

CAPÍTULO 4
Metodología

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

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

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

CAPÍTULO 5
Desarrollo

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

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

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

CAPÍTULO 6
Experimentos

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

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

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

CAPÍTULO 7
Conclusiones

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

Bibliografía

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

[2] Departamento de Lenguajes y Sistemas Informáticos. *Fuentes de Información*. Universitat Politècnica de València, versión 2, s.f.

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

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, segunda edición, 2006.

[5] National Eye Institute. Retinitis pigmentaria | National Eye Institute. Consultado el 13 de mayo de 2025. Disponible en https://www.nei.nih.gov/espanol/aprenda-sobre-la-salud-ocular/enfermedades-y-afecciones-de-los-ojos/retinitis-pigmentaria.

[6] Gil, G. A., Checa, F. L., Borrego, S., Chaparro-Hernández, P., Rueda, T. R., y Sanchez, J. Estudio de la variabilidad clínica y la heterogeneidad genética en la retinitis pigmentosa. Dialnet, 1994. Consultado el 13 de mayo de 2025. Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=6768158.

[7] Stone, E. M. Genetic Testing for Inherited Eye Disease. *Archives of Ophthalmology*, 125(2):205, 2007. https://doi.org/10.1001/archopht.125.2.205.

[8] Hanany, M., Rivolta, C., & Sharon, D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. *Proceedings of the National Academy of Sciences*, 117(5):2710–2716, 2020. https://doi.org/10.1073/pnas.1913179117.

[9] De Castro Miró, M. *Diagnóstico genético de las distrofias hereditarias de retina mediante secuenciación masiva de nueva generación (NGS)*. Tesis doctoral, Universitat de València, 2017. Repositorio UV. https://roderic.uv.es/handle/10550/61138.

[10] 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/.

[11] 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.

[12] 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.

[13] 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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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-

[19] F. Zhang, W. Gu, M. E. Hurles, i J. R. Lupski. Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481, 2009. https://doi.org/10.1146/annurev.genom.9.081307.164217.

[20] Roberto Togneri y Christopher J. S. deSilva. *Fundamentals of Information Theory and Coding Design*. CRC Press, Boca Raton, Florida, 2006.

APÉNDICE A

Configuració del sistema

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

A.1 Fase d'inicialització

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

A.2 Identificació de dispositius

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

APÉNDICE B
??? ????????????? ????

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