# Identificación y priorización de genes candidatos en enfermedades hereditarias: un estudio basado en redes génicas

José Carbonell Caballero

25 de agosto de 2012

**Resumen**

Esto será el abstract...

# Índice general


**1. Introducción** **2**
- 1.1. Biología molecular y computacional . . . 2
  - 1.1.1. Los inicios de la biología moculecular . . . 2
  - 1.1.2. Biología computacional . . . 3
- 1.2. Biología computacional en el estudio de enfermedades hereditarias 3
- 1.3. Secuenciación y la búsqueda de mutaciones . . . 5
  - 1.3.1. La secuenciación del genoma . . . 5
  - 1.3.2. Secuenciación por Sanger . . . 6
  - 1.3.3. Chips de ADN . . . 6
  - 1.3.4. Ultrasecuenciadores o técnicas de secuenciación masiva . . 6
- 1.4. Identificación y priorización de genes candidatos . . . 7
- 1.5. Metodologías de priorización en ausencia de genes conocidos . . . 7

**2. Material y métodos** **8**
- 2.1. Obtención de genes candidatos . . . 8
- 2.2. Metodología de priorización . . . 8
  - 2.2.1. Búsqueda de vecindarios compartidos entre familias . . . 8
  - 2.2.2. Medidas de distancia . . . 8
  - 2.2.3. Computo del estadístico de priorización . . . 8
  - 2.2.4. Diagrama general del método . . . 8
  - 2.2.5. Otras estrategias complementarias al método . . . 8
- 2.3. Validación . . . 8
  - 2.3.1. Simulaciones . . . 8
  - 2.3.2. Caso de uso . . . 8

**3. Resultados** **9**
- 3.1. Simulaciones . . . 9
- 3.2. Caso de uso . . . 9

**4. Conclusiones** **10**
- 4.1. Validez del método . . . 10
- 4.2. Aportación del método en ausencia de genes conocidos . . . 10
- 4.3. Limitaciones de la metodología y líneas futuras . . . 10

**5. Agradecimientos** **11**

**6. Material suplementario** **13**
- .1. First Appendix . . . 13

# 1 Introducción

## 1.1. Biología molecular y computacional

### 1.1.1. Los inicios de la biología moculecular

La biología molecular surge como una disciplina propia a raíz del descubrimiento de la doble hélice de ADN por parte de los investigadores James D. Watson y Francis Crick en 1953 [1], descubrimiento por el cual fueron galardonados, junto a Maurice Wilkins, con el premio nobel de medicina en 1962. Este hallazgo, que pasó desapercibido en un primer momento, fue el inicio de una serie de trabajos que permitieron describir como el ADN codifica en su interior las instrucciones necesarias para el funcionamiento de las células. Se descubrió que los aminoácidos se codificaban en grupos de 3 nucleótidos y que una secuencias de estos, formaba una proteína, la cual es la encargada directa de ejecutar las instrucciones contenidas en una determinada porción del ADN. A partir de ese momento, la biología molecular crece como una rama esencial de la biología y gracias a la influencia de otras ramas como la bioquímica, surge con un carácter marcadamente cuantitativo, dentro de una ciencia más acostumbrada a la descripción que a la medición.

En términos generales, la biología molecular se define como la parte de la biología encargada de estudiar los procesos celulares que ocurren a escala molecular. Estos procesos abarcan todos los mecanismos necesarios para que las células puedan funcionar correctamente, llevando a cabo todas sus funciones vitales con normalidad. Se trata de mecanismo generales que ocurren en todos los organismo conocidos, aunque por supuesto, con diferencias específicas en función del tipo celular.

La biología molecular nos permite comprender como funcionan las células y que mecanismos son esenciales para preservar la vida. Es de especial interés la composición de las diferentes formas de material genético (ARN o ADN) y los elementos moleculares que intervienen en su síntesis y regulación. 5 años después del descubrimiento de la doble hélice, Francis Crick formula lo que se conoce como dogma central de la biología molecular,. El dogma básicamente describe como el ADN de un gen contenido en el nucleo celular se trascribe a un ARN mensajero que sale del núcleo, y que posteriormente será traducido a una proteína que llevará a cabo la función del gen. A pesar de que con el tiempo el dogma central se ha desmitificado, ha servido para ilustrar de forma intuitiva como los genes son capaces de llevar a cabo su función en la célula.

### 1.1.2. Biología computacional

Desde los descubrimientos de Watson y Crick, la biología molecular ha crecido de la mano de un sinfín de tecnologías que le han permitido abordar de forma cuantitativa el estudio de los fenómenos moleculares. Desde el microscopio a los modernos secuenciadores, todos los descubrimientos han estado asentados sobre una base tecnológica muy fuerte, de la que han carecido otras áreas de la biología.

Una de las ramas tecnológicas que más ha evolucionado y aportado a la biología computacional en los últimos 30 años, ha sido la informática. Esta rama ha permitido procesar y almacenar de forma ordenada cantidades enormes de datos biológicos en tiempos de cómputo razonables. A la aplicación de la informática a la biología y su unión a otras disciplinas como la estadística o las matemáticas, se le denominó biología computacional.

La biología computacional, también denominada en ocasiones como bioinformática, es una disciplina en la que, a partir de la unión de modelos, algoritmos y computadores, es posible abordar la resolución de problemas biológicos. Los problemas más característicos afrontados en el pasado mediante la biología computacional han estado relacionados con el alineamiento de secuencias, lo que ha permitido entre otras cosas construir árboles filogenéticos que describen como evolucionan y se relacionan las distintas especies con sus ancestros comunes.

En los últimos años, el volumen de datos aportados por la biología ha crecido de forma considerable, lo que ha generado una dependencia informática, obligando a la incorporación de procedimientos y herramientas informáticas que permiten automatizar ciertos procesos casi cotidianos.

Entre los logros más importantes de la biología molecular y computacional destaca la secuenciación del primer genoma humano, tarea que fue resuelta en más de 13 años y en la que se invirtieron más de 280.000 millones de dólares por parte de algunos gobiernos y entidades privadas. La identificación de la secuencia del genoma humano y la localización de los genes ha tenido una gran relevancia en los ámbitos de la biomedicina y la genética clínica ya que aporta un visión genómica global que facilitó la interpretación de las bases moleculares de algunas enfermedades hereditarias.

## 1.2. Biología computacional en el estudio de enfermedades hereditarias

Las enfermedades de origen genético, son síndromes donde un error en la maquinaria de síntesis de proteínas provoca un comportamiento anómalo de las células y como consecuencia, la aparición de una enfermedad. Típicamente, su origen viene determinado por una mutación o variación estructural nociva en uno o varios genes, provocando un daño en cadena.

Cuando la enfermedad está causada por un único gen, decimos que es monogénica o mendeliana (en honor a Gregor Mendel). Ejemplos de enfermedades monogénicas son la fibrosis quística (causada por el gen *CFTR*), la enfermedad Huntington (causada por el gen HTT), o la Hemofilia de tipo A (causada por el gen F8). Por el contrario, cuando la enfermedad está causada por la combinación

de varios genes mutados, entonces decimos que es multigénica. Las enfermedades multigénicas son mucha más frecuentes que las monogénicas, y engloban la mayoría de las enfermedades crónicas como la hipertensión, la obesidad, o síndromes complejos como el alzheimer o la esquizofrénia.

Cuando la enfermedad es de carácter familiar, entonces decimos que es hereditaria. En este caso, la mutación o mutaciones nocivas se segregan de padres a hijos. Si un individuo acaba desarrollando una enfermedad de la que es portador vendrá determinado por lo que conocemos como modelo de herencia de la enfermedad.

Los humanos disponemos un genóma redundante compuesto por 23 pares de cromosomas, lo que significa a efectos prácticos (y a excepción de los cromosomas sexuales) que disponemos de dos copias de cada gen, donde una copia vendrá proporcionada por vía paterna y la otra por vía materna durante la meiosis. Esta duplicidad tiene efectos beneficiosos sobre el individuo y la especie, ya que genera (por combinación) biodiversidad y nos hace más robustos a enfermedades. Sin embargo, puede tener efectos funcionales ya que, aunque ambas copias deberían a ser prácticamente iguales, al provenir de parentales distintos cada copia puede disponer de mutaciones o alteraciones estructurales propias. Cuando una mutación está sólo en una copia, entonces decimos que se presenta en heterocigosis, sin embargo cuando la mutación es compartida por ambas, es decir, ambos parentales disponían de ella, entonces decimos que está en homocigosis. Cuando una enfermedad hereditaria muestra una herencia de tipo recesivo, significa que será necesario tener una mutación nociva en ambas copias para acabar desarrollando la enfermedad, lo que generalmente estará asociado a una pérdida de función del gen, ya que ninguna de sus proteínas seran funcionales. Sin embargo, cuando la herencia de la enfermedad es de tipo dominante, entonces significa que una sola copia mutada será suficiente para producir la enfermedad. En este caso, será irrelevante disponer de una copia intacta del gen, ya que generalmente la proteína mutada dispondrá de una nueva función que resultará nociva para la célula.

Comprender como funcionan los distintos modelos de enfermedad resulta absolutamente fundamental, ya que nos permite entender las bases moleculares de las enfermedades, y por consiguiente, avanzar en su prevención y cura. Además, nos permite ser mucho más eficientes a la hora de buscar que genes pueden estar implicados en una enfermedad con un origen genético total o parcialmente desconocido. El problema se complica cuando las variaciones genómicas no son responsables totalmente de la enfermedad. Entonces decimos que las condiciones ambientales (cómo los hábitos alimenticiones o el clima) explican un porcentaje significativo del riesgo a padecer la enfermedad. Existen enfermedades como el cáncer de pulmón, donde el ambiente (ser fumador) puede resultar totalmente decisivo en la proliferación de la enfermedad, y otras como las enfermedades monogénicas, donde presentar ciertas variaciones estructurales asegura una penetrancia casi total.

Biología computacional en la búsqueda de enfermedades***

Es conocido que no todas las mutaciones en genes van a producir los mismos efectos. Hay mutaciones que ni siquiera provocan un cambio de aminoácido en la secuencia proteíca, y otras que directamente impiden la síntesis de la proteina.

Mutaciones sinónimas/no sinónimas, No todas las mutaciones causan los mismos síntomas o la misma intensidad de la enfermedad ***

1000 genomas: mutaciones en población sana ***

Panorama mundial /penetrancia/ ***

Dentro del estudio de enfermedades hereditarias, es de especial interés el estudio de enfermedades raras. Las enfermedades raras, o también conocidas como huérfanas, son aquellas que tienen una baja incidencia en la población (inferior a 5 de cada 10.000 individuos), y por su condición de baja prevalencia, son sometidas a menores inversiones por parte de capital público o privado. Existen más de 7000 enfermedades raras descritas por la OMS y se estima que afectan al 7 % de la población mundial (sólo en España afectan a más de 3 millones de personas). Este tipo de síndromes se benefician claramente de las estrategías de secuenciación de genoma completo, ya que de otra forma, serían necesarios costosos estudios previos que focalizaran el origen de la enfermedad sobre un grupo reducido de genes candidatos.

## 1.3. Secuenciación de ADN

### 1.3.1. La secuenciación del genoma

La secuenciación del ADN nos permite conocer el orden específico de los nucleótidos en el genoma de un individuo, lo que posteriormente proporciona la identificación y localización de los genes. Comparando la secuencia obtenida con una secuencia de referencia es posible determinar aquellas diferencias o variaciones estructurales que presenta un individuo y que podrían ser susceptibles de haber causado o causar una enfermedad en el futuro. La construcción de un genoma de referencia con el que comparar ya supone de por sí un reto tecnológico importante que ha sido llevado a cabo en la última década. Además, el genoma de referencia debería constituir una plantilla sana con la que comparar, y definir la normalidad no es algo sencillo. Actualmente existe un consorcio internacional [2] encargado de construir y almacenar la versión oficial del genoma humano, actualizándolo de forma periódica.

A lo largo de los últimos 50 años han surgido multitud de método de secuenciación que han permitido conocer con fiabilidad la secuencia de nucleótidos de la que se compone un fragmento de material genético. Todos los métodos han tenido tasas de error cuantificables y parámetros como su complejidad, facilidad de ejecución, velocidad de secuenciación o coste económico han sido claves para su evolución y adopción por parte de los investigadores. Cabe tener en cuenta que todos los métodos de secuenciación han venido acompañados de herramientas y métodos estadísticos de análisis propios para su análisis, y que cada nueva generación de secuenciadores ha requerido un reaprendizaje de las nuevas herramientas de trabajo por parte de los investigadores. Este hecho ha tenido especial incidencia en los últimos años ya que cada tecnología ha dado lugar a una batería de herramientas informáticas de análisis y de algoritmos encargados de resolver problemas específicos de cada tecnología.

A continuación, se enumeran tres de las metodologías de secuenciación más empleados por la biología molecular.

### 1.3.2. Secuenciación por Sanger

El método de secuenciación más empleado en las últimas décadas ha sido el denominado método Sanger, en honor a su creador Frederick Sanger. Sanger, el cual es una de las 4 personas que ha recibido 2 veces un premio nobel en su vida, fue capaz de demostrar que las proteínas tienen estructura específica y que esta es fundamental para su función. Para ello, consiguió en 1955 determinar la secuencia de aminoácidos de la insulina y desarrolló un método con el que obtuvo un perfil específico de su estructura. Este trabajo le permitió obtener su primer nobel de química en 1958. Años más tarde (en 1975) desarrolla formalmente su método de secuenciación [3], el cual, a partir de dideoxinucleótidos, un gel de agarosa y la aplicación de electroforesis fue capaz de obtener un patrón de bandas a partir del cual es posible deducir la secuencia subyacente de nucleótidos. Con este método Sanger secuenció al bacteriófago A4, convirtiéndose en el primer organismo cuyo genoma fue secuenciado de forma completa. Los trabajos de Sanger fueron fundamentales en la consecución de proyecto genoma humano y en otros ambiciosos proyectos de secuenciación posteriores, gracias a lo cual obtuvo su segundo nobel de química en 1980.

### 1.3.3. Chips de ADN

A pesar de los importantes avances obtenidos mediante la secuenciación por Sanger, han sido necesario evolucionar a formas más automáticas de secuenciación que dejaran atrás algunas limitaciones técnicas como la secuenciación de largas cadenas de ADN. Con una perspectiva más de biología de sistemas, se ha evolucionado hacia métodos de secuenciación que son capaces de obtener información simultanea de miles de genes. Una de las tecnologías surgidas en la última década han sido los chips de ADN, los cuales, gracias su bajo coste y altas prestaciones, se han extendido a la gran mayoría de estudios de secuenciación importantes. Se trata de matrices de sondas (o pocillos), donde cada elemento es capaz de medir información concreta acerca un gen. Los chips más populares han sido los de expresión génica. Estos, compuestos por pocillos que contiene la secuencia complementaria de cada gen, son capaces de medir la abundancia de ARN mensajeros y por tanto de la expresión de cada gen. Otros chips ampliamente extendidos, han sido los chips de genotipado. En este caso, el chip se encarga de testar la presencia de una cantidad enorme de polimorfismo en la secuencia de los genes. Si bien, no ha sido una tecnología a partir de la cual obtener la secuencia precisa de nucleótidos de cada gen, si han servido de forma eficiente y barata en la determinación de las características básicas de los genes.

### 1.3.4. Ultrasecuenciadores o técnicas de secuenciación masiva

Los avances de la última década han permitido el desarrollo de técnicas de secuenciación másiva a un coste muy bajo. Se trata de tecnologías que permiten procesar y secuenciar simultáneamente millones de fragmentos de ADN. Son las llamadas tecnologías de ultrasecuenciación o secuenciación másiva, y son capaces de secuenciar un genoma completo en aproximadamente una semana, con un coste inferior a 20.000 dolares. Lógicamente, estas técnicas han revolucionado

la genómica computacional, dirigiendo los estudios hacia una perspectiva más genómica que génica.

A pesar de sus numerosas prestaciones, los técnicas de secuenciación másiva también muestran algunas limitaciones importantes. La más destacada reside en la longitud máxima que son capaces de secuenciar. Actualmente, los secuenciadores son capaces de obtener fragmentos de entre 50 a 500 nucleótidos, lo cual, comparado con técnicas clásicas como la secuenciación por Sanger, resulta muy pobre. Por esta razón los secuenciadores precisan un paso de computado adicional denominado mapeo. Este paso básicamente se encarga de volver a posicionar (o localizar) cada fragmento secuenciado sobre genoma de referencia. Algunos parámetros como la longitud del fragmento serán claves para la fiabilidad del mapeo, ya que, teniendo en cuenta que la secuencia del genóma dispone de sólo 4 nucleótidos diferentes, a menor tamaño, más probable será encontrar la misma porción en varios puntos del genoma. Otra limitación importante reside en la tasa de error a la hora de ir obtenido la secuencia de nucleótidos, la cual es superior a otras técnicas como la secuenciación por Sanger. Por esta razón, es habitual que después de la fase de análisis estadístico, se recurra al método Sanger para validar las mutaciones obtenidas mediante los ultrasecuenciadores.

A pesar de todo, los ultrasecuenciadores constituye la tecnología más prometedora y, gracias a su enfoque genómico, resultando muy eficientes en el estudio de aquellas enfermedades donde los genes causantes de la enfermedad son desconocidos, y un enfoque más centrado en genes individuales sería totalmente infructuoso.

## 1.4. Identificación y priorización de genes candidatos

Los genes constituyen únicamente el 5 % del genoma, el resto, antaño considerado como ADN basura, tiene una función todavía desconocida, que en algunas ocasiones estará relacionada con aspectos más estructurales que funcionales en la molécula de ADN. A su vez, los genes tienen en su interior cierto grado de estrucutura. Poseen porciones que serán codificadas directamente a proteínas, llamadas regiones codificantes o exones, pero también poseen regiones relacionadas con su regulación, regiones empleadas durante la transcripción y otras regiones denominadas intrones situadas alrededor de los exones.

Debido a la estructura y funcionamiento del genoma, es mucho más probable encontrar una mutación causante de enfermedad dentro de una zona codificante, que en zonas intergénicas, u otras zonas interiores al gen, menos sensibles a un cambio de nucleótido. Este factor produce que en la práctica no se realice una secuenciación orientada a genoma completo, por el contrario, se recurre a kits especiales de secuenciación que cubren únicamente las regiones codificantes de los genes. Esta reducción en el área de búsqueda tiene ciertas limitaciones, ya que en algunas ocasiones la variación estructural no aparecerá en la zona codificante, sino en zonas destinadas a la regulación del gen. A cambio, la porción de ADN estudiado a sólo un 1 % del total, lo cual aporta grandes beneficios para toda la fase de procesamiento, almacenamiento y análisis que ocurre después de la secuenciación.

El protocolo de selección de genes candidatos se realiza a partir de un estudio caso/control. En este, se seleccionan una serie de individuos que comparten el mismo síndrome (habitualmente causado por la misma mutación o gen), acompañado de un grupo de individuos control que nos permitirá filtrar todas aquellas mutaciones potencialmente nocivas que por estar presentes en población sana no deberían estar relacionadas con síndrome. La forma inicial de obtener los candidatos consiste en seleccionar aquellas mutaciones o genes mutados con una prevalencia significativamente mayor en los casos que en los controles. Debido al reducido tamaño muestral de los estudios, y ha algunos errores de secuenciación, la mutación causante de la enfermedad será seleccionando junto a un grupo de mutaciones adicionales, que no podran ser filtradas. Los genes que contienen a las mutaciones seleccionadas, serán considerados como genes candidatos a la enfermedad.

El tamaño muestral y otras parámetros de error relacionados con la secuenciación y los pasos de cómputo adicional, tendrán una influencia directa sobre el número de genes candidatos.

identificación de mutaciones***

Modelos animales***

Selección de genes candidatos***

priorización basada en funciones biológicas***

comparación con genes conocidos***

## 1.5. Metodologías de priorización en ausencia de genes conocidos

# 2 Material y métodos

## 2.1. Obtención de genes candidatos

## 2.2. Metodología de priorización

### 2.2.1. Búsqueda de vecindarios compartidos entre familias

### 2.2.2. Medidas de distancia

### 2.2.3. Computo del estadístico de priorización

### 2.2.4. Diagrama general del método

### 2.2.5. Otras estrategias complementarias al método

**Estimación del rol general del gen**

**Información funcional**

## 2.3. Validación

### 2.3.1. Simulaciones

### 2.3.2. Caso de uso

# 3 Resultados

## 3.1. Simulaciones

## 3.2. Caso de uso

# 4 Conclusiones

## 4.1. Validez del método

## 4.2. Aportación del método en ausencia de genes conocidos

## 4.3. Limitaciones de la metodología y líneas futuras

# 5 Agradecimientos

# Bibliografía

[1] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[2] Genome reference consortium. `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/`.

[3] F Sanger, S Nicklen, and AR Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of The United States Of America*, 74(12):5463–5467, 1977.

# 6 Material suplementario

## .1. First Appendix