TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2016 201

Antonio Martín Navarro

MITOCLASS.1, un predictor de patogenicidad para mutaciones no sinónimas en los polipéptidos codificados por el mtDNA humano

Departamento
Bioquímica y Biología Molecular y Celular

Director/es
Mayordomo Cámara, Elvira
Ruiz Pesini, Eduardo

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

MITOCLASS.1, UN PREDICTOR DE PATOGENICIDAD PARA MUTACIONES NO SINÓNIMAS EN LOS POLIPÉPTIDOS CODIFICADOS POR EL MTDNA HUMANO

Autor

Antonio Martín Navarro

Director/es

Mayordomo Cámara, Elvira
Ruiz Pesini, Eduardo

UNIVERSIDAD DE ZARAGOZA

Bioquímica y Biología Molecular y Celular

2016

MITOCLASS.1. Un predictor de patogenicidad para mutaciones no sinónimas en los polipéptidos codificados por el mtDNA humano

Antonio Martín Navarro

"Jamás fui tan consciente de lo lejos que me encontraba de mi meta como cuando estuve tan cerca de ella".

Gattaca (Película, 1997)

INDICE

1. Resumen

La secuenciación del mtDNA de pacientes con enfermedades mitocondriales está revelando muchas nuevas mutaciones no sinónimas por lo que se hace necesario priorizar qué substituciones son interesantes de someter a estudios de confirmación de patogenicidad. Para este cribado previo de substituciones resulta útil el uso de programas predictores que permitan un adecuado filtrado de aquellas mutaciones que, a priori, cabría esperar que mostraran un fenotipo patológico. Sin embargo, estos programas todavía no han demostrado poseer una sensibilidad y especificidad adecuadas para su utilización con este objetivo.

Durante los últimos años han surgido diferentes métodos predictores de la patogenicidad de mutaciones no sinónimas pero ninguno ha sido específicamente diseñado para la predicción de variantes de polipéptidos codificados por el mtDNA. Además, todavía no existe una base de datos correctamente depurada de substituciones patológicas en el mtDNA por lo que no puede llevarse a cabo una adecuada evaluación de los programas clasificadores disponibles.

Fruto de nuestra investigación, hemos logrado desarrollar un programa clasificador basado en aprendizaje automático Naive Bayes de mutaciones patológicas exclusivo para variantes no sinónimas del mtDNA. El entrenamiento y validación de nuestro modelo ha sido realizado con 2835 substituciones de aminoácidos neutras y patológicas previamente revisadas siguiendo unos criterios de patogenicidad definidos igualmente en nuestro laboratorio y presentados en este trabajo.

Cada mutación está descrita por un conjunto de tres atributos basados tanto en conservación evolutiva en organismos eucariotas para cada posición como en la posibilidad de que dicha posición muestre coevolución con otras del mismo polipéptido. También hemos incluido un novedoso atributo basado en el análisis de la conservación evolutiva de cada uno de los veinte aminoácidos en cada uno de los tres dominios de los polipéptidos codificados por el mtDNA (dominio intermembrana, transmembrana y matriz). Para ello, fue necesaria una caracterización previa de las posiciones presentes en cada dominio tampoco realizada hasta la fecha.

Previamente, evaluamos las prestaciones de tres predictores ampliamente utilizados (Polyphen-2, Provean y Mutpred) utilizando la base de datos de 2835 substituciones elaborada por nuestro grupo de investigación. El predictor Polyphen-2

resultó ser el más adecuado para su uso como test de cribado por su buena sensibilidad. Sin embargo, el número de falsos positivos obtenido fue elevado y además, un porcentaje minoritario de mutaciones no pudieron ser clasificadas por dicho predictor.

Nuestro método, Mitoclass.1 ha mostrado una sensibilidad mejorada sobre Polyphen-2 para un conjunto de 1100 mutaciones utilizadas para la validación del test. Mitoclass.1 también ha reflejado una mejora en la especificidad y además no presenta variantes sin clasificar.

Adicionalmente, hemos incluido en este trabajo los resultados predictivos para el conjunto completo de variantes posibles (24201) de los trece polipéptidos codificados por el mtDNA. Un porcentaje importante de las variantes (68,8 %) ha resultado ser susceptible de resultar patológico y su confirmación por estudios adicionales sería interesante.

Así pues, Mitoclass.1 permite una mejor selección de variantes no sinónimas posiblemente patogénicas. Sus mejores prestaciones en relación a otros predictores tiene su origen en la cuidadosa selección de los atributos discriminadores utilizados en la etapa de entrenamiento del clasificador, así como en el hecho de haber utilizado una base de datos de substituciones no sinónimas correctamente depurada y exclusiva de genes codificantes del mtDNA humano.

Mitoclass.1 es un clasificador que podría optimizarse en el futuro con la publicación de nuevas variantes no sinónimas en el mtDNA humano en bases de datos como GenBank. Por ello, resultaría muy interesante actualizar periódicamente los valores numéricos de los discriminadores de Mitoclass.1 utilizando dicha información y revalidando el test.

2. Abreviaturas

- AUC: Área bajo la curva
- BLAST: Basic Local Alignment Search Tool
- CI: Índice de conservación
- cMI: Información mutua acumulada
- ETC: Cadena transportadora de electrones
- HGMD: Human Genome Mutation Database
- IM: Dominio intermembrana
- M: Dominio matriz
- MAFFT: Multiple Alignment using Fast Fourier Transform
- MCC: Coeficiente de correlación de Matthews
- mdmv.1: Mitochondrial DNA missense variants versión 1
- MI: Información mutua
- MISTIC: Mutual information server to infer coevolution
- mtDNA: DNA mitocondrial
- NCBI: National Center for Biotechnology Information
- NUMT: DNA mitocondrial nuclear
- OXPHOS: Fosforilación oxidativa
- PDB: Protein Data Bank
- PERL: Practical Extraction and Report Language
- PPV: Valor predictivo positivo
- rCRS: Secuencia de Cambridge revisada
- RefSeq: Base de datos "Reference Sequence"
- RNA: Acido ribonucléico
- ROC: Característica operativa del receptor
- ROS: Especies reactivas de oxígeno
- rRNA: RNA ribosómico
- TM: Dominio transmembrana
- tRNA: RNA de transferencia
- WEKA:Waikato Environment for Knowledge Analysis

3. Introducción

3.1. El sistema de fosforilación oxidativa

La fosforilación oxidativa (OXPHOS) es un proceso metabólico que utiliza energía liberada por la oxidación de nutrientes para producir adenosina trifosfato (ATP). Consta de dos etapas: en la primera, la energía libre generada mediante reacciones químicas redox en varios complejos multiproteicos, conocidos en su conjunto como cadena de transporte de electrones, se emplea para producir un gradiente electroquímico de protones a través de una membrana asociada en un proceso llamado quimiosmosis. La cadena respiratoria está formada por tres complejos de proteínas principales (complejo I, III, IV), y varios complejos "auxiliares", utilizando una variedad de donantes y aceptores de electrones. Los tres complejos se asocian en supercomplejos para canalizar los electrones, haciendo más eficiente el proceso. Los polipéptidos integrantes de estos complejos son codificados mayoritariamente por el DNA nuclear. A pesar de ello, existen trece polipéptidos que son codificados por el DNA mitocondrial (mtDNA) en la especie humana: siete (p.MT-ND1-ND6 y ND4L) del complejo I (NADH:ubiquinona oxidoreductasa), uno (p.MT-CYB) del complejo III (ubiquinol:citocromo c oxidorreductasa), tres (p.MT-CO1-CO3) del complejo IV (citocromo c oxidasa) y dos (p.MT-ATP6,ATP8) del complejo V (ATP sintasa).

3.2. El DNA mitocondrial humano

El DNA mitocondrial humano (mtDNA) es una molécula circular de aproximadamente 16,5 kilobases (Anderson et al., 1981; Andrews et al., 1999). Existen muchas copias de este genoma por mitocondria y cientos de mitocondrias por célula. Se transmite a través de herencia materna y codifica 37 genes, 2 RNA ribosómicos (12 S y 16S rRNA), 22 RNA de transferencia (tRNA) y 13 polipéptidos, todos ellos componentes del sistema de fosforilación oxidativa (OXPHOS).

El mtDNA se localiza en la matriz mitocondrial, cerca de la membrana interna, lugar en el que se encuentra ubicada la cadena transportadora de electrones (ETC). La ETC es la fuente principal de especies reactivas de oxígeno (del inglés ROS), tales

como superóxido y peróxido de hidrógeno que pueden provocar daño celular y facilitar mutaciones en el mtDNA.

El proceso generador de mutaciones inducido por ROS está favorecido por el hecho de que el genoma mitocondrial no está cubierto de histonas, a pesar de que existan diferentes proteínas como las mtTFA dedicadas a la organización del genoma. Además, el mtDNA puede replicarse más de una vez durante el ciclo celular y durante la replicación, largos segmentos de mtDNA pueden permanecer como cadenas simples permitiendo la formación de estructuras secundarias que afecten al correcto funcionamiento de la DNA polimerasa promoviendo la aparición de nuevas mutaciones en el mtDNA.

A esto hay que añadir que los sistemas de reparación del mtDNA no son tan abundantes como los presentes en el DNA nuclear (Druzhyna et al., 2008), haciendo que el mtDNA pueda acumular mutaciones a una velocidad mayor que el DNA nuclear (Montoya et al., 2009).

Debido a que los genes del mtDNA codificantes de polipéptidos representan el 70 % del mtDNA humano, la mayor parte de estas mutaciones afectan a genes codificantes de proteínas. Algunas de estas mutaciones son responsables de enfermedades muy graves, pero muchas otras no muestran efectos fenotípicos importantes.

Hay que tener en cuenta que las substituciones de aminoácidos pueden ser debidas a transiciones o transversiones en el DNA. Las substituciones simples de un nucleótido por otro se denominan transiciones o transversiones, según provoquen el cambio de purina por purina ó pirimidina por pirimidina (transiciones), o el cambio de una purina por pirimidina ó viceversa (transversiones). Así, algunas mutaciones ocurren más frecuentemente que otras y se ha observado que en el mtDNA animal existe un exceso de transiciones sobre transversiones (Keller et al., 2007).

3.3. Enfermedades genéticas del DNA mitocondrial

Las manifestaciones clínicas de estas enfermedades son muy variadas (Munnich et al., 1996) y entre las más comunes encontramos: demencia, trastornos motores, intolerancia al ejercicio, accidentes cerebrovasculares, convulsiones, ptosis, oftalmoplejía, retinopatía pigmentaria, atrofia óptica, ceguera, sordera, cardiomiopatía, disfunciones hepáticas y pancreáticas, diabetes, falta de crecimiento, anemia

sideroblástica, pseudobstrucción intestinal, nefropatías, estatura corta, acidosis metabólica, y otras más secundarias.

En general, son trastornos multisistémicos que afectan fundamentalmente a los tejidos y órganos que más dependen de la energía mitocondrial (sistema nervioso central, músculo cardíaco y esquelético, riñones y sistema endocrino). Sin embargo, dado que las mitocondrias están presentes en todos los tejidos, otros muchos órganos pueden estar implicados en estos síndromes tan heterogéneos. De hecho, una de las pistas que conduce a la sospecha de enfermedad mitocondrial es la implicación de muchos órganos diferentes. No obstante, algunas enfermedades claramente mitocondriales no presentan estos caracteres tan típicos, especialmente en pacientes en edad pediátrica. Todo lo comentado demuestra la gran dificultad existente actualmente para diagnosticar una enfermedad mitocondrial y más aún para vincularla a una mutacion del mtDNA y no del DNA nuclear.

Dada la heterogeneidad de las manifestaciones clínicas, morfológicas y bioquímicas presentes, su clasificación se basa en las características moleculares y genéticas de las mutaciones (Figura 1). Así, las enfermedades mitocondriales pueden dividirse en dos grandes grupos:

a) enfermedades asociadas a mutaciones puntuales.

b) enfermedades debidas a reorganizaciones del mtDNA por inserciones y/o deleciones.

Muchas mutaciones de estos genes codificantes de proteínas son mutaciones puntuales (que afectan a un solo nucleótido) no sinónimas (que provocan una substitución de un aminoácido por otro).

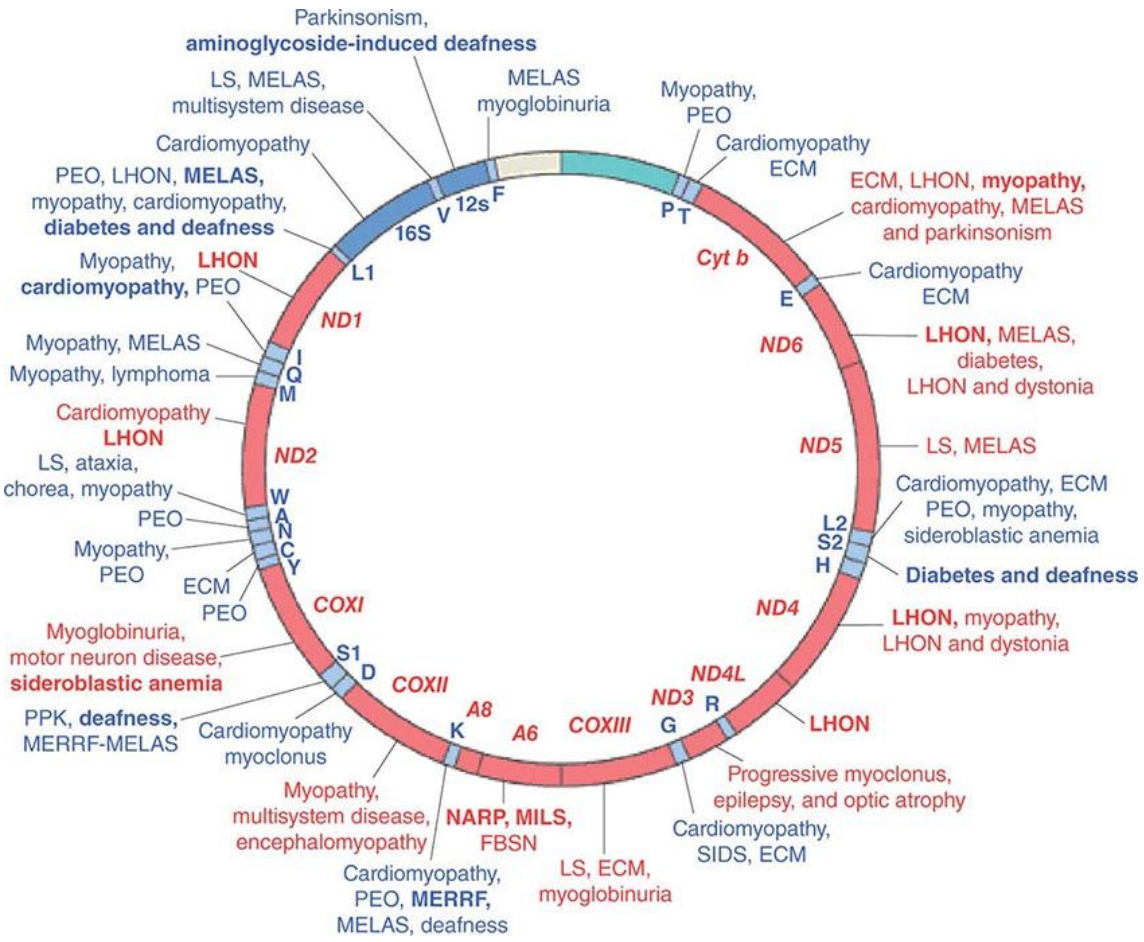

Figura 1. Molécula de DNA mitocondrial humano y enfermedades vinculadas a mutaciones en sus genes. ((DiMauro and Schon, 2003))

3.4. Criterios de patogenicidad de variantes no sinónimas del mtDNA

No es tarea fácil diferenciar mutaciones potencialmente patológicas de otras de carácter neutro y durante los últimos años se han propuesto diferentes criterios para conseguir discriminar unas de otras (Elson et al., 2012; Montoya et al., 2009).

Actualmente, todavía no existe un consenso definitivo sobre qué criterios requiere cumplir una mutación en el mtDNA de un paciente para que sea clasificada como la causa de la enfermedad. En un trabajo publicado por Dimauro y Schon en 2001 se describe el criterio canónico de patogenicidad para confirmar si una mutación puntual es la causa de la enfermedad. Primero, la mutación debe ser heteroplásmica (en la misma célula existen diferentes genomas de mtDNA). Segundo, el mtDNA mutado debe segregar con un defecto bioquímico en tejidos clínicamente relevantes y preferiblemente con signos de enfermedad severa en miembros de la familia procedentes de herencia materna. Tercero, la variante debe pertenecer a una posición

conservada a través de la evolución y no tiene que estar presente en un grupo de población control étnicamente relacionado (DiMauro and Schon, 2001).

Sin embargo, algunas mutaciones patogénicas identificadas desde entonces no cumplen este criterio canónico. De hecho, un número cada vez mayor de mutaciones patogénicas homoplásmicas están siendo identificadas y por lo tanto, no cumplen con la primera condición. Además, no afectan siempre a posiciones muy conservadas de los polipéptidos, por lo que la conservación por sí sola no es suficiente para determinar si un cambio es patológico (McFarland et al., 2004a; Tuppen et al., 2008).

Existen varios estudios en los que se tratan de establecer los mejores criterios para determinar si un cambio en el mtDNA es un buen candidato para ser clasificado como patológico, tanto para genes tRNA (McFarland et al., 2004b; Yarham et al., 2011) como para genes codificantes de proteínas (Mitchell et al., 2006). Estos estudios establecen una puntuación de patogenicidad que incluye varios parámetros como la presencia de heteroplasmia, segregación dentro de la familia, defectos bioquímicos, conservación interespecífica y estudios funcionales.

Otra dificultad en la clasificación de variantes patológicas ha sido la presencia de pseudogenes mitocondriales en el genoma nuclear (NUMTs) que fueron incorrectamente amplificados en el pasado en pacientes con posibles patologías mitocondriales haciendo vincular de forma anómala la patología a genes codificados por el mtDNA (Yao et al., 2008).

Por otro lado, si una variante define un haplogrupo o un subhaplogrupo es improbable que dicho cambio sea la causa principal de una enfermedad mitocondrial. A pesar de eso, dichas variantes podrían por ejemplo modular la penetrancia de otra mutación y originar así un fenotipo patológico. Por ello, el conocimiento insuficiente en la actualidad en relación a la filogenia mitocondrial es también una fuente de incorrecciones a la hora de definir la patogenicidad de un cambio (Bandelt et al., 2005, 2007).

En el trabajo de Montoya (Montoya et al., 2009) se establecen algunas consideraciones importantes sobre los criterios a seguir. Por ejemplo, la población control debe estar correctamente definida, siendo necesario un número elevado de controles tanto geográficamente como genéticamente relacionados con el potencial paciente. Por otro lado, la mutación debe encontrarse en las puntas del árbol filogenético del mtDNA, aunque es cierto que a veces mutaciones patológicas han aparecido en las ramas internas del árbol debido a cambios compensatorios en otros

lugares del genoma. Además, el genoma del mtDNA debe ser completamente secuenciado para descartar otras mutaciones que pudieran ser igualmente candidatas de resultar patológicas o incluso la presencia de dos mutaciones patológicas dentro del mismo paciente. Otro aspecto a considerar es el hecho de que las mutaciones patológicas suelen ser recesivas pero han aparecido casos de mutaciones con efecto dominante, por lo que también debe ser tenido en cuenta este mecanismo. Es importante igualmente tener presente la interacción con otras posiciones de la molécula o de moléculas vecinas y asumir que la función de muchos polipéptidos codificados por el mtDNA así como de diferentes dominios proteicos es todavía desconocida.

Finalmente, también es necesario recordar que los estudios funcionales como el uso de cíbridos son buenos modelos pero tampoco están exentos de problemas. La conclusión es que las patologías mitocondriales son complejas y que no se puede ser estricto en la aplicación de los criterios de patogenicidad. A pesar de ello, estos criterios son una herramienta interesante para considerar si una nueva mutación pendiente de clasificar puede ser o no realmente patológica.

Una consecuencia inherente al hecho de que muchas mutaciones definidas actualmente como patológicas probablemente no lo sean es la presencia de errores en las bases de datos. Es muy posible que bases de datos de referencia contengan mutaciones mal clasificadas que no sean realmente patológicas o neutras. Usar dichos listados de mutaciones para, por ejemplo, entrenar a un predictor bioinformático que clasificara una mutación todavía no catalogada provocaría sesgos importantes. Es por ello necesario en la actualidad un proceso de curación exhaustiva de dichas bases de datos.

3.5. Utilidad de los predictores bioinformáticos de patogenicidad

La caracterización funcional de una mutación no sinónima en el mtDNA es una prueba irremplazable hoy en día para determinar su efecto fenotípico y su potencial patogenicidad. Estos estudios funcionales como el uso de cíbridos transmitocondriales permiten no sólo confirmar el efecto patogénico, sino también determinar la severidad de la mutación y poder predecir su evolución, estudiar su mecanismo molecular, proponer posibles terapias futuras y analizar la relación estructura-función de la proteína (residuos catalíticos, regulatorios, estructurales, etc...). Desafortunadamente, el análisis funcional no siempre es posible debido tanto a su coste económico como al tiempo de

trabajo requerido. Además, ya que las enfermedades mitocondriales pueden también ser originadas por mutaciones en el DNA nuclear (Goldstein et al., 2013), la búsqueda de técnicas que permitan priorizar el estudio funcional de una determinada mutación potencialmente patológica es crucial. Por esta razón, los predictores bioinformáticos son muy útiles debido a su rapidez y mínimo coste económico. En los últimos años se han convertido en un instrumento valioso para el estudio de mutaciones no sinónimas no aparecidas previamente en pacientes con una posible enfermedad mitocondrial y poder decidir así si es interesante su posterior estudio funcional confirmatorio. De todos modos, es importante tener presente que estos programas predictores muestran todavía limitaciones y que es necesario contar con más evidencias para confirmar si las variantes predichas son realmente patológicas.

Actualmente, con las mejoras en rapidez y bajada de costes de la secuenciación de un genoma completo humano, el mayor foco de atención de los estudios genéticos está centrado en una pequeña porción (1 %) del genoma encargada de la codificación de proteínas (conocida como "exoma") (Stenson et al., 2009). En el exoma humano se encuentran la mayor parte de variantes asociadas a enfermedades y se han identificado más de 20.000 polimorfismos en el exoma de personas sanas (Bamshad et al., 2011). La mitad de estos cambios son mutaciones no sinónimas con cambios de aminoácido en las proteínas. El gran desafío de los investigadores consiste actualmente en la interpretación de toda esta información y en el desarrollo computacional de estrategias para identificar a esa minoría de variantes que son realmente susceptibles de producir patologías (Ohanian et al., 2012).

3.6. Tipos de predictores disponibles

Podemos distinguir entre tres tipos principales de predictores:

a) El primer grupo engloba a los programas basados en el análisis de la secuencia de las proteínas y en su conservación evolutiva. A favor, estos métodos aprovechan el hecho de que las variantes patológicas suelen estar localizadas en posiciones conservadas. En contra, el hecho de que estos programas son muy sensibles a los alineamientos múltiples de secuencias homólogas a la humana proporcionados por el usuario siendo la predicción muy dependiente del número de secuencias alineadas y su distancia evolutiva. Algunos importantes predictores como PROVEAN (Choi et al., 2012), SIFT (Kumar et al., 2009) , Align-GVGD (Tavtigian et al., 2006), Mutation assessor (Reva et

al., 2011), Panther (Brunham et al., 2005) o MAPP (Stone and Sidow, 2005) se basan en estos métodos.

b) El segundo grupo engloba los predictores basados en el análisis de la secuencia y la estructura de la proteína. A favor de estos métodos se encuentra el considerar que otros aspectos además de la conservación evolutiva puede ayudar en la clasificación de las variantes. En contra, el hecho de que todavía no existe un número elevado de estructuras cristalinas publicadas. Además de ello, algunos de estos algoritmos generan informes con mucha información estructural que necesitan ser interpretados por el usuario para poder decidir si la mutación es o no patológica. Esto requiere un buen conocimiento de algunos de esos parámetros para evitar que la información sea confusa o interpretada de forma errónea. Este grupo contiene a uno de los predictores de referencia, Polyphen-2 (Adzhubei et al., 2010) y a otros como SNPEffect (De Baets et al., 2012) y diferentes métodos que analizan la estabilidad de la proteína: FoldX (Schymkowitz et al., 2005), PopMusic (Dehouck et al., 2011), MuPro (Cheng et al., 2006) y SDM (Worth et al., 2011).

c) El tercer grupo engloba a los métodos basados en aprendizaje automático supervisado (supervised-learning method), es decir, métodos que utilizan la información completa disponible sobre mutaciones ya clasificadas para inferir la patogenicidad de nuevas mutaciones. Estos algoritmos incluyen técnicas como los "neural networks", las máquinas de vectores de soporte ("support vector machines o SVM), los "random forest" y los clasificadores "naive Bayes". El argumento a favor de estos métodos es la posibilidad de combinar un amplio grupo de parámetros discriminadores predictivos para clasificar más adecuadamente las variantes, analizando factores que no pueden ser capturados usando información de secuencia o estructural. Por contra, requieren bases de datos con números grandes de mutaciones clasificadas para entrenarse y además, muchas veces, estas bases de datos usadas por los programas no están correctamente depuradas encontrándose mutaciones patológicas que en realidad no lo son o mutaciones neutras que son en realidad patológicas. En este tercer grupo encontramos un amplio y creciente número de programas. Entre los más importantes, podemos destacar: PMut (Ferrer-Costa et al., 2004), SNAP (Bromberg et al., 2008), PhD-SNP (Capriotti et al., 2006), SNPs&GO (Calabrese et al., 2009), MutPred (Li et al., 2009), Hansa (Acharya and Nagarajaram, 2012) y MutationTaster (Schwarz et al., 2010).

3.7. Selección de la base de datos de variantes humanas

Un parámetro importante a tener en cuenta es el conjunto de mutaciones que van a ser utilizadas para la etapa tanto de validación del predictor como de entrenamiento del mismo en caso de que se trate de un clasificador de aprendizaje automático. En este último supuesto, el programa necesita entrenarse para discernir entre lo que es una mutación patológica y una mutación neutra.

3.8. Selección de los parámetros discriminadores

Por último, es también clave la selección de los parámetros discriminadores con los que el predictor va a evaluar si una mutación no clasificada es o no patológica.

En el trabajo de Thursberg (Thusberg et al., 2011) se concluye que, de nueve predictores evaluados, no existe ninguno que pueda ser catalogado como el más adecuado siendo por tanto necesario el desarrollo de nuevos métodos más fiables. Así pues, el usuario debe considerar qué aspectos son los más importantes en su trabajo ya que existen diferencias en los atributos discriminadores de unos y otros programas.

El índice de conservación (CI), es un criterio comúnmente utilizado para la determinación de patogenicidad en los polipéptidos codificados por el mtDNA (DiMauro and Schon, 2001; Montoya et al., 2009). Para desarrollar este análisis, el único requerimiento es la ejecución de un alineamiento múltiple de secuencias de polipéptidos ortólogos al humano. Sin embargo, algunos trabajos han demostrado que la exactitud de estas herramientas en posiciones ultraconservadas es baja y las predicciones generadas pueden confundir a los investigadores en el posterior estudio clínico o experimental de mutaciones (Castellana and Mazza, 2013; Gnad et al., 2013; Martelotto et al., 2014). Por ello, el análisis de la conservación por si solo no es suficiente para discriminar substituciones patológicas y se hace imprescindible evaluar nuevos atributos.

4. Hipótesis de trabajo y objetivos

Nuestra hipótesis de trabajo es muy sencilla, aunque nuestro objetivo general es muy ambicioso. Dado que los programas para la predicción de patogenicidad de mutaciones no sinónimas más ampliamente utilizados se han creado fundamentelmente a partir de proteínas codificadas en los cromosomas nucleares y con función en distintos compartimentos celulares, pensamos que podemos mejorar su poder predictivo si generamos una herramienta bioinformática de predicción de patogenicidad basada en proteínas codificadas en el mtDNA y con función exclusiva en la membrana interna mitocondrial.

Así pues, nuestro objetivo principal es la consecución de un programa de predicción de patogenicidad de mutaciones no sinónimas en el mtDNA que sea realmente eficiente.

Para ello será necesario plantear varios objetivos secundarios previos:

a) Establecer una base de datos compuesta exclusivamente por variantes tanto patológicas como neutras de polipéptidos codificados por el mtDNA para entrenar correctamente al predictor. La elaboración de nuestra base de datos supone revisar el conjunto de variantes definidas previamente como patológicas en la literatura para, de acuerdo con unos criterios de patogenicidad definidos en nuestro laboratorio, eliminar las que carezcan de evidencias suficientes para serlo.

b) Elegir la combinación de atributos discriminadores que ofrezca mayor fiabilidad y siempre mejorada respecto de otros predictores ya disponibles. En nuestro caso, combinaremos discriminadores típicos como el estudio de la conservación con otros más novedosos, como el análisis del grado de coevolución entre posiciones de un mismo polipéptido. También evaluaremos la posibilidad de que una misma substitución pueda tener diferente efecto si se encuentra en el dominio transmembrana, en el espacio matriz o en el dominio intermembranoso del polipéptido ya que todos los polipéptidos codificados por el mtDNA son proteínas integrales de membrana con estos tres dominios zonales bien diferenciados.

c) Seleccionar el algoritmo de clasificación más adecuado que permita discriminar mutaciones neutras y patológicas.

5. Métodos

5.1. Consideraciones previas sobre el diseño del predictor

El aprendizaje automático es una rama de la inteligencia artificial (Narayanan et al., 2002) cuyo objetivo es desarrollar técnicas que permitan a un ordenador “aprender”. Se trata de crear un programa capaz de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Nuestro predictor está basado en un algoritmo de aprendizaje supervisado denominado "Naive Bayes" o clasificador bayesiano ingenuo (Figura 2). Se trata de un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. El clasificador recibe el apelativo de "ingenuo" a causa de estas simplificaciones, resumidas en una hipótesis de independencia entre las variables predictoras (Larrañaga et al., 2006). En este método, dado un conjunto de muestras de entrenamiento, podemos etiquetar las clases a las que pertenecen (mutaciones patológicas y neutras en nuestro trabajo) y entrenar al predictor para construir un modelo que prediga la clase de una nueva muestra. Una buena separación entre las clases permitirá una clasificación correcta.

El correcto diseño de un clasificador de patogenicidad de mutaciones no sinónimas requiere cumplir varias premisas. La primera de ellas es disponer de una base de datos de mutaciones no sinónimas tanto patológicas como neutras bien caracterizada, sin presencia de falsos positivos ni falsos negativos. A continuación, seleccionar un conjunto de atributos discriminadores que permitan diferenciar con la máxima confianza posible entre ambos grupos de mutaciones.

Una vez establecido el modelo final del predictor, procederemos a la validación del mismo. En esta fase, se debe tener en cuenta que no es adecuado utilizar instancias pertenecientes al subconjunto de mutaciones empleado en la etapa de entrenamiento ya que los resultados sobre ellas no representarían el verdadero poder de generalización del predictor y falsearían los resultados de la validación (Vihinen, 2012). Por ello, la base de datos de mutaciones requiere ser dividida antes de entrenar al predictor en dos subconjuntos denominados “training dataset” (grupo de entrenamiento) y “validation dataset” (grupo de validación).

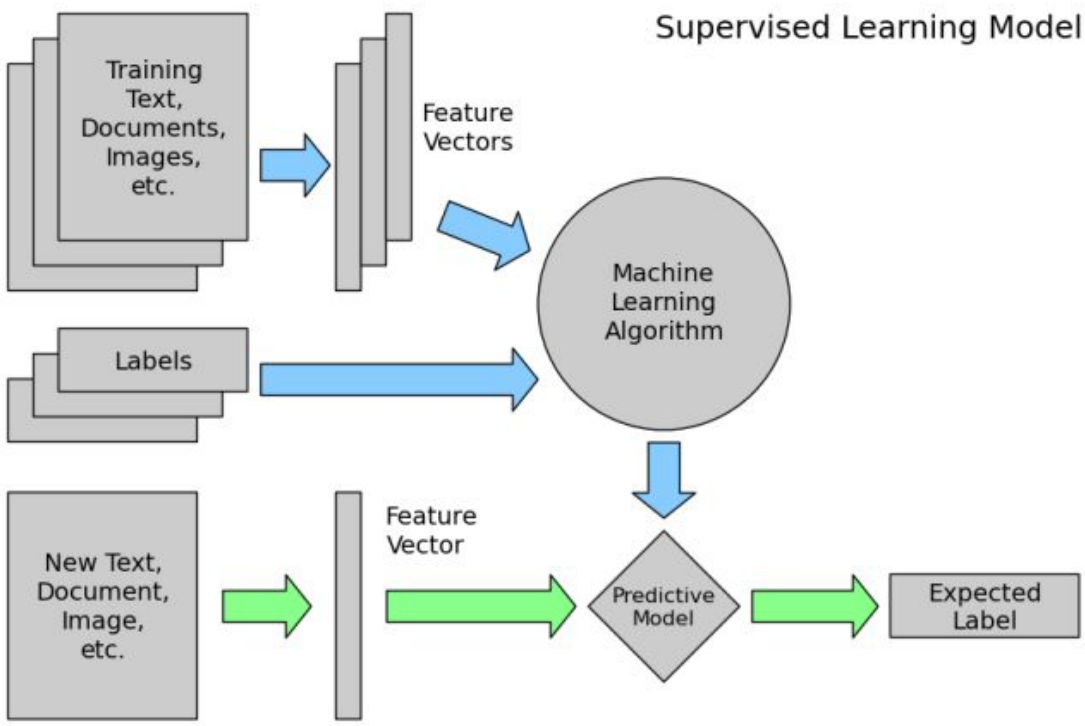

Figura 2. Diagrama general de funcionamiento de un clasificador basado en aprendizaje supervisado. (extraído de http://morganpolotan.me/)

5.2. Criterios de patogenicidad para identificación de variantes patológicas

Las mutaciones no sinónimas son aquellas en la que se produce un cambio en un nucleótido de la cadena de DNA que produce un cambio en el codón correspondiente que trae como consecuencia la incorporación de un aminoácido diferente.

Sin embargo, ciertas mutaciones no sinónimas clasificadas en la sección de Mitomap "mtDNA Mutations with Reports of Disease-Associations" sólo cumplen algunos de los criterios establecidos por la bibliografía para considerar la mutación como patológica (Montoya et al., 2009) siendo por tanto su clasificación como tal, dudosa. Para salvar este inconveniente, en nuestro estudio solamente hemos clasificado como patológicas aquellas substituciones que han sido previamente asociadas a una posible enfermedad mitocondrial y cumplen, por lo menos, uno de los dos criterios definidos a continuación.

Tras la aplicación de estos criterios sobre las variantes no sinónimas presentes en la sección de Mitomap “mtDNA Mutations with Reports of Disease-Associations”, hemos generado la base de datos mdmv.1 incluyendo finalmente 57 variantes patológicas y 2778 variantes neutras. Estas variantes no patológicas incluyen aquellas

mutaciones de la sección de Mitomap denominada "mtDNA Mutations with Reports of Disease-Associations" que no cumplían los criterios que se explican a continuación para considerarlas como patológicas, así como las mutaciones presentes en la sección "mtDNA Variants" que, en principio, engloba todo el conjunto de variantes publicadas con fenotipos neutros.

5.2.1. Criterio primero: Confirmación del origen genómico mitocondrial de la patología a través de estudios funcionales

Estos estudios son una evidencia de peso a la hora de designar patogenicidad a una determinada mutación (Mitchell et al., 2006). En nuestro trabajo hemos analizado para cada mutación patológica publicada si dicha patogenicidad fue confirmada a través de estudios funcionales basados en cíbridos transmitocondriales o en fibras individuales. En ambas técnicas, diferentes genotipos de mtDNA se asocian a un mismo fondo nuclear y ambiental. Así, las diferencias funcionales entre cíbridos o entre fibras musculares individuales con diferente carga mutacional pueden justificarse por cambios en el genotipo del mtDNA.

Las funciones mitocondriales están controladas tanto por el mtDNA como por el DNA nuclear. Por ello, es difícil identificar cuál de los dos genomas es el responsable de un defecto en la mitocondria. El uso de cíbridos es una herramienta muy útil para dilucidar si el genoma del mtDNA es el responsable de un defecto. Para ello, se comparan mitocondrias procedentes de diferentes fuentes en un entorno nuclear definido. Los cíbridos (Figura 3) se construyen fusionando células anucleadas portadoras o no de una determinada mutación mitocondrial con células en las cuales el mtDNA endógeno ha sido eliminado (células ρ^0). De esta manera, se pueden analizar las consecuencias de alteraciones en el mtDNA a nivel celular excluyendo la influencia de mutaciones en el DNA nuclear (Vithayathil et al., 2012).

Cuando no todas las moléculas de mtDNA de una célula están mutadas, se dice que el individuo es heteroplásmico. Si el porcentaje de mutación es muy elevado, esa célula o fibra muscular puede presentar un déficit en la actividad de alguno o todos los complejos con subunidades codificadas en el mtDNA. Por tanto, en el análisis de fibras individuales se realiza una tinción histoquímica para enzimas oxidativas como por ejemplo la detección de actividad de la citocromo c oxidasa para asociar el grado de tinción con el porcentaje de mutación. Las fibras musculares o células afectadas tendrán

cargas de mutación mayores que las fibras o células no afectadas. Dado que las células se obtienen del mismo tejido de un mismo paciente, el fondo genético y el ambiente es el mismo (Johnson et al., 1993).

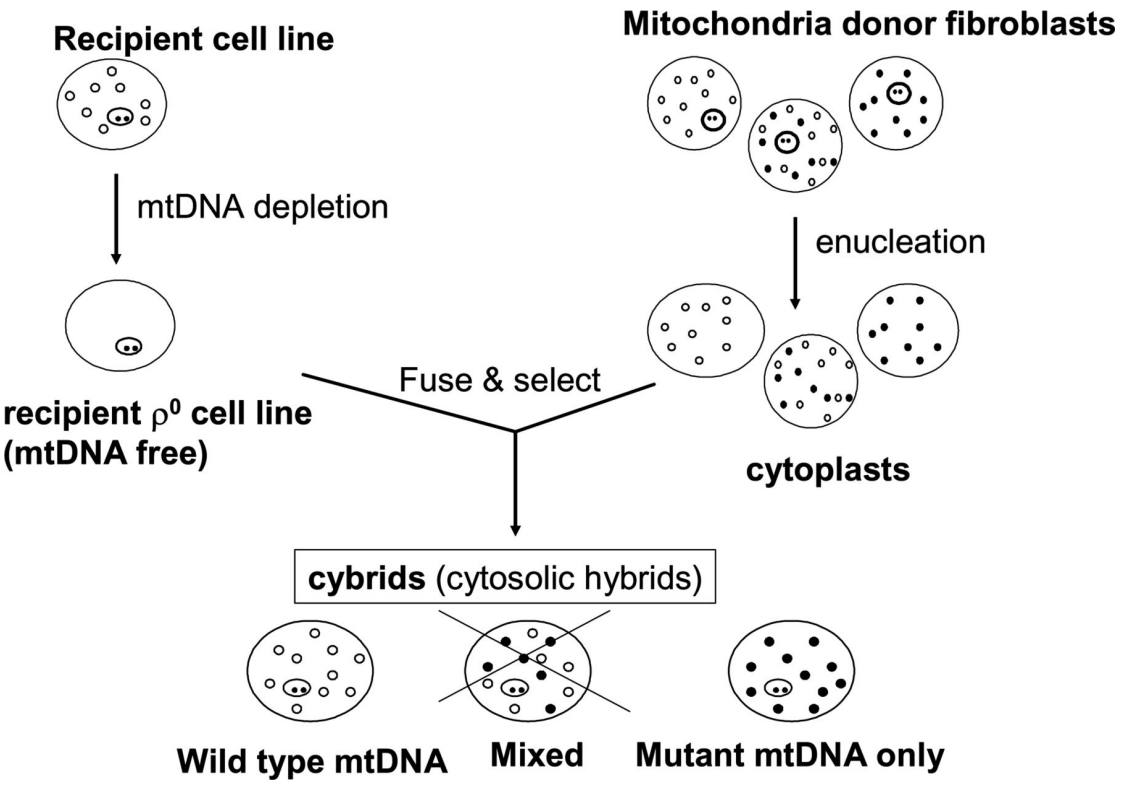

Figura 3. Proceso de generación de cíbridos transmitocondriales. (Maechler and de Andrade, 2006)

5.2.2. Criterio segundo: Análisis de la frecuencia polimórfica en humanos de la variante potencialmente patológica

Hemos analizado esta frecuencia considerando el conjunto de secuencias humanas del polipéptido publicadas en GenBank. Las enfermedades raras afectan a un número limitado de individuos, definido como no más de uno por cada 2000 individuos de la Unión Europea (Schieppati et al., 2008). Las enfermedades mitocondriales pertenecen a este grupo de patologías y están presentes en alrededor de uno de cada 10000 adultos (Chinnery, 1993). Teniendo en cuenta estas consideraciones, hemos clasificado como patológicas en nuestro trabajo únicamente aquellas mutaciones presentes en más de un pedigrí con pacientes enfermos de mitocondriopatía pero ausentes en la población control o presentes en la población control pero con una frecuencia muy baja (≤ 0.1 ‰).

En todos los casos, para apoyar este criterio, hemos verificado que estos cambios no aparecen en las ramas internas del árbol filogenético mitocondrial humano. Cuando una nueva mutación surge en el mtDNA, una nueva rama queda establecida en el árbol filogenético mitocondrial. Así, las mutaciones más recientes deberán aparecer en las puntas de las ramas del árbol y aquellas que sobrevivan a selección purificadora (mecanismo de selección natural que elimina las mutaciones deletéreas) quedarán fijadas en las ramas más internas. Por ello, las mutaciones patológicas no permanecerán en la población por mucho tiempo (Fan et al., 2008). Del mismo modo, aunque se han encontrado casos de mutaciones patológicas presentes en ramas internas (Cheng et al., 2006) debido a mutaciones compensatorias en otras posiciones del genoma (Wang et al., 2006), será improbable encontrar dichas mutaciones patológicas en las ramas internas del árbol (Figura 4).

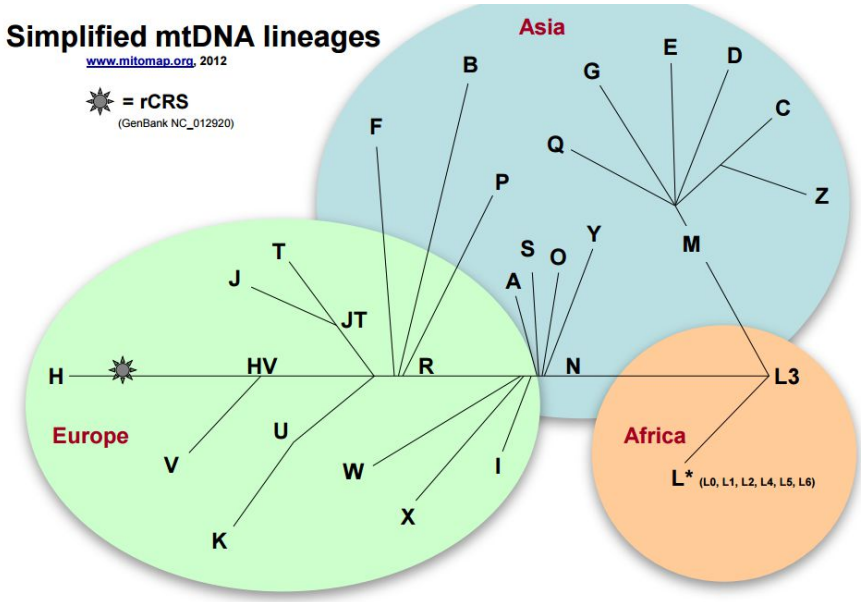

Figura 4. Árbol filogenético mitocondrial humano (extraído de www.mitomap.org)

5.3. Caracterización bioinformática de los dominios de los polipéptidos humanos codificados en el mtDNA

Todos los polipéptidos codificados por el mtDNA son proteínas integrales de membrana (Figura 5) con tres dominios bien diferenciados: un dominio en la matriz, otro en el espacio intermembranoso y el último en la membrana interna de la

mitocondria (dominio transmembrana). El ambiente bioquímico de cada una de estas tres regiones es diferente. El espacio intermembranoso tiene una alta concentración de protones (pH ácido) como resultado del bombeo de los mismos por los complejos enzimáticos de la cadena respiratoria. En la matriz mitocondrial el pH es más básico. Por otro lado los aminoácidos insertados en el interior de la membrana interna suelen ser residuos hidrofóbicos elegidos para interactuar con los lípidos de la bicapa lipídica mientras que los ubicados tanto en el espacio intermembranoso como en la matriz son habitualmente más hidrofílicos. Por ello, una misma substitución en cada uno de dichos dominios podría tener efectos funcionales diferentes (Stefl et al., 2013).

Desafortunadamente, no existen todavía estructuras cristalinas disponibles para estos polipéptidos humanos. El hecho de que uno de los discriminadores de Mitoclass.1 requiera distinguir el dominio en el que se produce una substitución hace necesario un trabajo previo de caracterización de dominios (localizar las posiciones de la molécula presentes en cada uno de ellos). Esta caracterización se ha realizado utilizando proteínas ortólogas (aquellas secuencias homólogas que se han separado evolutivamente por un proceso de especiación y que son altamente similares por proceder de un ancestro común) de otros organismos cuya estructura cristalina si se encuentra disponible en la base de datos Protein Data Bank (PDB) (Tabla 1). El PDB es una base de datos compuesta por estructuras tridimensionales de proteínas y ácidos nucleicos. Estos datos, generalmente obtenidos mediante cristalografía de rayos X o resonancia magnética nuclear, están bajo dominio público y pueden ser usados de forma gratuita (Berman et al., 2000).

Polipéptido humano	Código PDB	Organismo con estructura cristalina
p.MT-ND1	4HEA (chain H)	*Thermus thermophilus*
p.MT-ND2	4HEA (chain N)	*Thermus thermophilus*
p.MT-ND3	4HEA (chain A)	*Thermus thermophilus*
p.MT-ND4	4HEA (chain M)	*Thermus thermophilus*
p.MT-ND4L	4HEA (chain K)	*Thermus thermophilus*
p.MT-ND5	4HEA (chain L)	*Thermus thermophilus*
p.MT-ND6	4HEA (chain J)	*Thermus Thermophilus*
p.MT-CYB	1QCR (chain C)	*Bos taurus*
p.MT-CO1	1OCC (chain A)	*Bos taurus*
p.MT-CO2	1OCC (chain B)	*Bos taurus*
p.MT-CO3	1OCC (chain C)	*Bos taurus*
p.MT-ATP6	1C17 (chain M)	*Escherichia coli*
p.MT-ATP8	----------	----------

Tabla 1. Proteínas homólogas con estructura cristalina conocida utilizadas en nuestro estudio de caracterización de dominios de los polipéptidos humanos codificados por el mtDNA.

Para el caso del polipéptido p.MT-ATP8 no es posible encontrar una secuencia ortóloga de origen bacteriano debido a que se trata de un polipéptido supernumerario, no presente en especies procariotas. Muchas de estas proteínas conocidas como supernumerarias o "accesorias" están presentes en todos los eucariotas mientras que otras son más específicas de un linaje concreto. Probablemente esto se debe a que en las primeras etapas de la evolución de los seres vivos eucariotas, los complejos I, III, IV y V reclutaron un gran número de proteínas codificadas por el núcleo para el sistema de fosforilación oxidativa mitocondrial (Sluis et al., 2015). Además, todavía no existen estructuras cristalinas para dicho polipéptido en otras especies eucariotas. Por ello, hemos utilizado para su caracterización la información contenida en trabajos previos (Hong and Pedersen, 2004).

Estos dominios han sido anotados con ayuda del programa Cn3D 4.3.1, que es una herramienta de visualización de estructuras cristalinas (Wang et al., 2000) que permite correlacionar estructura con secuencias y alineamientos. De esta manera, hemos correlacionado la estructura cristalina de un polipéptido ortólogo al humano con el alineamiento entre dicho ortólogo y la secuencia humana, para extraer así las posiciones iniciales y finales ocupadas teóricamente por la molécula humana en cada hélice transmembrana. Posteriormente, utilizando el programa de visualización de estructuras

químicas Jmol (Herráez, 2006) con las estructuras cristalinas de los complejos respiratorios completos publicadas en otras especies eucariotas, hemos podido determinar en qué dominio se encuentra el extremo N amino-terminal y el extremo C carboxilo-terminal de cada polipéptido. Combinando así las dos informaciones (localización de los dominios transmembrana y localización de los extremos amino y carboxilo-terminal de las moléculas), ha sido posible determinar también las posiciones de cada polipéptido presentes en los dominios intermembranoso y matriz.

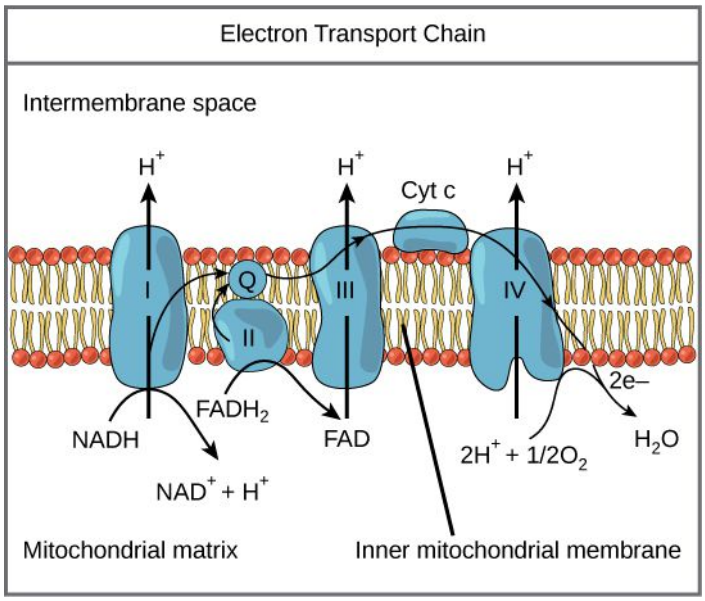

Figura 5. Esquema de la cadena transportadora de electrones mitocondrial y sus cuatro complejos respiratorios.

5.4. Cálculo del índice de conservación en especies eucariotas

El índice de conservación evolutiva en eucariotas (CI) es un parámetro utilizado por los atributos discriminadores del predictor Mitoclass.1. Por ello, describimos a continuación el procedimiento seguido para su cálculo.

Una serie de comandos en lenguaje Perl, denominados Bioperl Eutilities (Stajich et al., 2002) permiten descargar desde la base de datos GenBank (Benson et al., 2015) todos los polipéptidos codificados por el mtDNA, ortólogos a los humanos, procedentes de organismos presentes en la base de datos RefSeq del NCBI (alrededor de 5000 especies en febrero de 2015). La colección Reference Sequence (RefSeq) contiene un conjunto de secuencias de genomas, transcritos y proteínas bien anotadas y no redundantes. Hemos elegido dicha base de datos por poseer una única secuencia de

referencia para cada organismo, de forma que permite generar un conjunto de secuencias ortólogas para cada una de las humanas evitando más de una secuencia por especie y manteniendo un grado de curación y fiabilidad adecuado.

Los alineamientos de secuencias múltiples se han realizado con el programa MAFFT v.7.147b (--auto option) (Katoh and Standley, 2013). MAFFT ha sido elegido por ser uno de los métodos más rápidos y de mejor exactitud de entre las herramientas de alineamiento múltiple actualmente disponibles. Además, su exactitud ha sido probada frente a otros alineadores con buenos resultados (Pervez et al., 2014) y el programa ha sido empleado en proyectos importantes de reconocido prestigio como Pfam (base de datos de familias de proteínas) (Finn et al., 2014).

Un fichero "fasta" es un formato de fichero informático basado en texto, utilizado para representar secuencias que permite incluir nombres de secuencias y comentarios que preceden a las secuencias en sí. Cada fichero fasta generado por MAFFT con el alineamiento de las secuencias ortólogas a un polipéptido codificado por el mtDNA humano se ha utilizado como argumento de entrada de un script escrito en lenguaje Perl y Awk (lenguaje de programación diseñado para procesar datos basados en texto). Este script calcula inicialmente la frecuencia absoluta de los aminoácidos presentes en cada posición y a continuación el CI, definido como la frecuencia relativa del aminoácido presente en el polipéptido humano incluido en la base de datos RefSeq procedente del genoma humano de referencia NC_012920. Los valores del CI se representan en forma de porcentaje.

5.5. Atributos discriminadores

5.5.1. Discriminador 1: CI + cMI en Eucariotas

Este discriminador es la suma de dos parámetros: CI y cMI.

A veces cuando una mutación tiene lugar en una determinada posición de una proteína, hay una o varias mutaciones en posiciones diferentes de la proteína que covarían. A este mecanismo se le denomina coevolución (Figura 6). Por ello, una substitución en una posición que coevoluciona con otra de esa misma proteína puede resultar patológica si no ocurre una mutación compensatoria en el residuo con el que coevoluciona.

Para determinar este parámetro hemos utilizado el programa MISTIC (mutual information server to infer coevolution) (Simonetti et al., 2013). Este programa está disponible online y permite un análisis completo de las redes (networks) de información mutua en familias de proteínas. Para ello hay que aportar al programa los ficheros de alineamientos múltiples de proteínas ortólogas a las humanas. Esta puntuación de información mutua (MI) (Martin et al., 2005) generada por MISTIC puede usarse para estimar el grado de coevolución entre dos posiciones de una familia de proteínas ortólogas. La información mutua o transinformación de dos variables aleatorias es una cantidad que mide la dependencia mutua de las dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria, X, debido al conocimiento del valor de otra variable aleatoria Y.

Un parámetro derivado del MI denominado cumulative Mutual Information (cMI) también generado por MISTIC ha sido el utilizado en nuestro trabajo. El cMI suma el valor del MI de cada posición con todas las demás del polipéptido. Por tanto, valores altos de cMI indican residuos con alta posibilidad de coevolución (sin profundizar en los residuos con los que coevoluciona). Debido a que los valores de cMI presentan diferencias cuantitativas entre polipéptidos debido al diferente tamaño de las moléculas (ya que el número de residuos que coevolucionan con uno dado será a priori mayor si la molécula contiene más aminoácidos), hemos considerado normalizar los valores a una escala de 0 a 100 % para cada polipéptido aplicando como líneas de base los valores mínimo y máximo de cMI de cada uno de ellos. El resultado es un cMI relativo que será sumado con el CI para definir el valor numérico del discriminador 1.

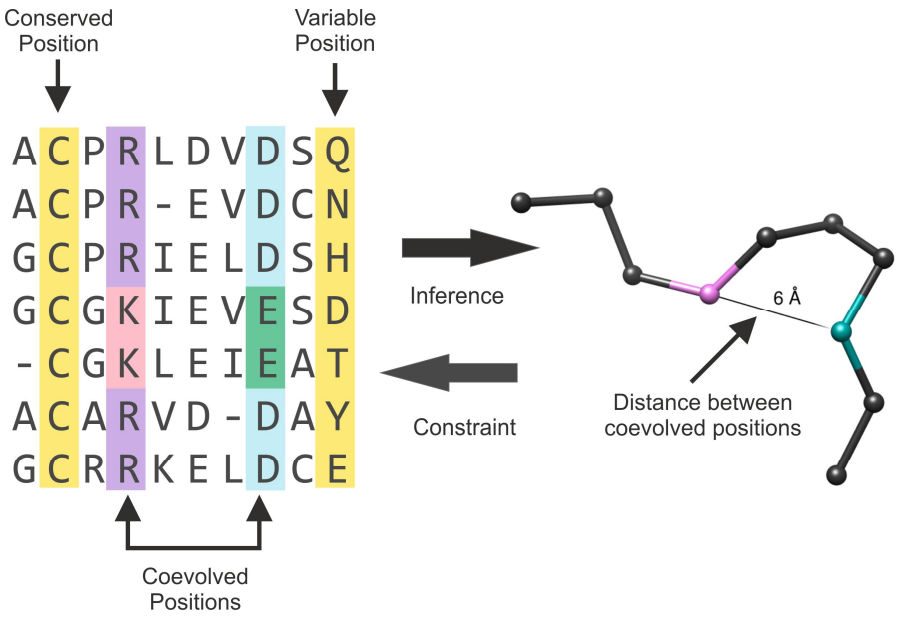

Figura 6. Representación de un alineamiento múltiple y de la estructura de una proteína del alineamiento mostrando las posiciones conservadas y variables, así como las que coevolucionan (figura adaptada de Marks et al., 2011).

5.5.2. Discriminador 2: Frecuencia de aparición del aminoácido mutante en cada posición de los polipéptidos

Este parámetro se obtiene junto con el CI de cada posición como resultado de los scripts ya explicados para el discriminador 1. Para cada posición, la conservación del aminoácido salvaje (presente en la secuencia de referencia NC_012920.1) en todo el set de ortólogos analizados queda definida como el CI mientras que la presencia del resto de aminoácidos aparecidos en dicha posición correspondería con la frecuencia relativa de los aminoácidos mutantes. En este caso las posiciones ocupadas por gaps de alineamiento han sido tenidas en cuenta para el cálculo de las frecuencias relativas.

5.5.3. Discriminador 3: Frecuencia de aparición de aminoácidos mutantes para cada tipo de aminoácido en un mismo dominio

Dado que los trece polipéptidos codificados por el mtDNA humano son proteínas integrales de membrana con tres dominios bien diferenciados (transmembrana, intermembranoso y matriz), hemos generado una tabla de frecuencias relativas para cada posible variante en cada uno de los tres dominios. Esta frecuencia relativa se basa en los valores de CI y frecuencia absoluta de aparición de cada variante por posición en

eucariotas y está basada en el alineamiento múltiple de secuencias ortólogas obtenido para los trece polipéptidos. La secuencia genómica humana de referencia (rCRS, NC_012920.1) fue seleccionada para determinar el aminoácido de referencia en cada posición para cada gen. Finalmente, todas las posiciones pertenecientes al mismo dominio fueron agrupadas para generar la tabla de variantes de los 20 tipos distintos de aminoácidos incluyendo también la frecuencia de cambio a un gap de alineamiento.

Para el cálculo del valor numérico del discriminador 3 se excluyó la frecuencia relativa de cada aminoácido consigo mismo (su conservación o CI en el dominio, mostrado en la diagonal de la tabla) y la frecuencia de aparición de gaps. De esta manera, queda repartido el 100 % del valor del discriminador para cada tipo de aminoácido entre las frecuencias de cambio a los 19 aminoácidos diferentes.

5.6. Método de aprendizaje automático elegido para Mitoclass.1

La base de datos completa mdmv.1 con 2835 variantes fue dividida de forma aleatoria en dos. El primer subgrupo se utilizó para constituir la base de datos de entrenamiento y estaba formada por el 60 % de mutaciones neutras y el 60 % de mutaciones patológicas (1735 mutaciones totales). El otro subgrupo se empleó para constituir la base de datos de validación con 1100 mutaciones.

Para la selección del clasificador más adecuado se efectúo una validación cruzada de 10 iteraciones sobre la base de datos de entrenamiento empleando diferentes clasificadores (Bayes, trees, funciones, etc...). En la validación cruzada de K iteraciones los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. La aplicación Weka 3.7.7 (Hall et al., 2009) fue utilizada para la ejecución de la validación cruzada con los diferentes algoritmos de clasificación evaluados. Weka (Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») es una plataforma de software para el aprendizaje automático y la minería de datos escrita en Java y desarrollada en la Universidad de Waikato. Weka es un software libre distribuido bajo la licencia GNU-GPL.

El clasificador Naive Bayes Simple fue el que logró mejores resultados y por ello, se utilizó para constituir el predictor Mitoclass.1.

5.6.1. Balanceo de datos de cada clase en la base de datos mdmv.1

Debido al diferente de número de mutaciones de cada clase (patológicas y neutras) que componen la base de datos de entrenamiento utilizamos el algoritmo SMOTE (Chawla et al., 2011) para sobrerrepresentar la clase minoritaria (patológica) y disponer de un número similar de mutaciones en ambas clases. Los parámetros elegidos en SMOTE fueron nearestneightbors=5 y randomseed=1. SMOTE crea muestras sintéticas de la clase minoritaria interpolando muestras cercanas a diferencia de otros métodos de sobrerrepresentación basados en crear directamente copias. El problema de las clases con distinto número de instancias está ampliamente documentado (Guo et al., 2008) dado que muchos problemas reales presentan esta característica. En ellos, casi siempre la clase menos representada suele ser la más importante, como en nuestro caso (mutaciones patológicas) y los clasificadores tratan de optimizar la precisión total en los datos de entrenamiento, lo que lleva a que intenten clasificar bien los datos mayoritarios. Esto genera una buena precisión final, pero a costa de obtener un número excesivo de falsos negativos sobre la clase menos representada (Wasikowski and Chen, 2010).

5.7. Predictores utilizados en la comparación con Mitoclass.1

5.7.1. Mutpred

Este predictor se basa también en aprendizaje automático, utilizando un clasificador del tipo "Random Forest". Mutpred analiza los resultados del predictor SIFT (Ng and Henikoff, 2003), así como una serie de discriminadores relacionados con la pérdida o ganancia de 14 propiedades estructurales o funcionales (ganancia en la predisposición a formar una hélice, pérdida de un lugar de fosforilación, etc...). Mutpred se ha entrenado con mutaciones patológicas de la base de datos Human Gene Mutation Database (Stenson et al., 2009) y con polimorfismos neutros de la base de datos Swiss-Prot (Boeckmann et al., 2003). En nuestro estudio hemos seleccionado como mutaciones patológicas aquellas con puntuación igual o mayor a 0,75 (Li et al., 2009)

tal y como recomiendan los desarrolladores del programa en su página web (http://mutpred.mutdb.org/about.html).

Los resultados predictivos de Mutpred utilizados en la comparación han sido los definidos en un trabajo previo publicado en el año 2011 (Pereira et al., 2011).

5.7.2. Polyphen-2 version 2.2.2

Este predictor utiliza una combinación de atributos basados en secuencia y estructura de las proteínas. El efecto de la mutación queda predicho a través de un clasificador "naive Bayes". Polyphen-2 genera una puntuación numérica y además, un resultado cualitativo, clasificando las mutaciones en tres grupos denominados "benign", "possibly damaging" y "probably damaging", de menos a más posiblemente patológico.

En nuestro estudio, para una comparación adecuada entre predictores, hemos considerado como patológicas tanto las predicciones "probably damaging" como las "possibly damaging" (Adzhubei et al., 2010).

5.7.3. Provean version 1.1.3

PROVEAN (Protein Variation Effect Analyzer) es un programa que predice si la substitución de un aminoácido tiene un impacto negativo en la función biológica de la proteína. En resumen, el predictor utiliza BLAST para descargar un conjunto de proteínas homólogas a una dada. Después, el programa CD-HIT realiza un agrupamiento de las secuencias que presentan un 75 % de identidad global en la secuencia. Los 30 grupos conteniendo a las secuencias más similares forman un grupo de secuencias de soporte (los autores lo denominan "supporting sequence set"). Este grupo sirve para generar la predicción. Para cada una de las secuencias del grupo de soporte se calcula una puntuación de alineamiento delta (delta alignment score). Estas puntuaciones se promedian dentro de cada grupo y entre los 30 grupos para definir la puntuación final (PROVEAN score). Si la puntuación es igual o menor a un punto de corte determinado (por defecto -2,5), la variante se predice como patológica. Si la puntuación está por encima del punto de corte, la variante se predice como neutra (Choi et al., 2012).

5.8. Evaluación del predictor

Para evaluar la calidad de la predicción de Mitoclass.1 hemos efectuado un test ciego sobre el previamente definido grupo de validación (constituido por todas las mutaciones de mdmv.1 excluyendo las mutaciones utilizadas en el grupo de entrenamiento). En nuestro trabajo hemos calculado la sensibilidad, la especificidad y el coeficiente de correlación de Mathews (MCC).

El MCC (Matthews, 1975) se utiliza ampliamente en aprendizaje automático para medir la calidad de las clasificaciones binarias o de dos clases. El coeficiente tiene en cuenta los positivos verdaderos y los falsos positivos y puede usarse incluso si las clases son de diferente tamaño. El MCC es, en esencia, un coeficiente de correlación entre las clasificaciones observadas y predichas. Su valor oscila entre -1 y +1. Un coeficiente de +1 representa una predicción perfecta. Un coeficiente de 0 indica una predicción similar a una predicción aleatoria, mientras que -1 representaría un total desacuerdo entre predicción y observación.

En las ecuaciones siguientes, TP, TN, FP y FN se refieren al número de positivos verdaderos, negativos verdaderos, falsos positivos y falsos negativos según su terminología inglesa respectivamente.

Sensibilidad (S) = TP/(TP+FN)

Especificidad (E) = TN/(TN+FP)

$$MCC = (TP \times TN) - (FP \times FN) / \sqrt{(TP+FP)\,(TP+FN)\,(TN+FN)\,(TN+FP)}$$

Para comparar visualmente al predictor con los otros programas se han generado las curvas ROC utilizando lenguaje R (paquete pROC) (Robin et al., 2011) y también se han calculado y comparado los valores AUC (area under the curve) de cada predictor.

Una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a (1 – especificidad) para un sistema clasificador binario según se varía el umbral de discriminación. El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda del espacio ROC, representando un 100 % de sensibilidad (ningún falso negativo) y un 100 % también de especificidad (ningún falso positivo). A este punto también se le llama una clasificación perfecta. Por el contrario, una clasificación totalmente aleatoria daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación, desde el extremo inferior izquierdo hasta la

esquina superior derecha. La diagonal divide el espacio ROC. Los puntos por encima de la diagonal representan los buenos resultados de clasificación (mejor que el azar). La curva ROC se puede usar para generar estadísticos que resumen la efectividad del clasificador. Uno de ellos es el área bajo la curva ROC, llamada comúnmente AUC.

Este índice se puede interpretar como la probabilidad de que un clasificador ordene o puntúe una instancia positiva elegida aleatoriamente más alta que una negativa. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria predictiva (Figura 7).

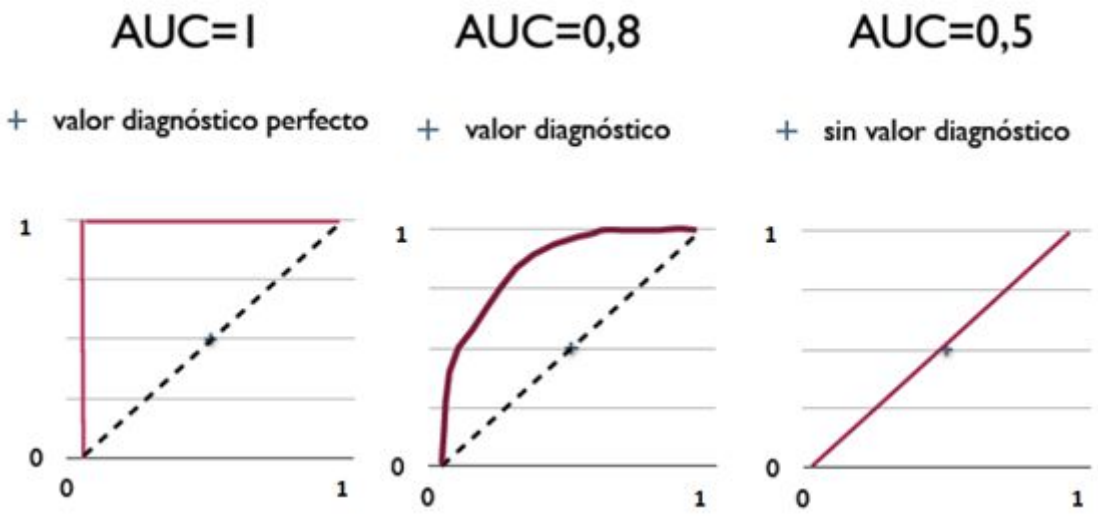

Figura 7. Ejemplos de valores diagnósticos de curvas ROC con diferentes valores AUC.

5.9. Análisis estadístico

El test estadístico no paramétrico Mann-Whitney-Wilcoxon se ha ejecutado usando lenguaje R para decidir si la distribución de las poblaciones de datos (por ejemplo, las patológicas y neutras de un determinado discriminador) es idéntica a un nivel de significación de $p \leq 0,05$.

Por otro lado, el coeficiente de correlación de Spearman se utilizó para la justificación de uso del índice de conservación (CI) como parámetro indicador de conservación interespecífica comparándolo con otros indicadores de conservación documentados.

5.10. Otros parámetros cuantificados

5.10.1. Cálculo del índice de conservación evolutivo (CI) de los polipéptidos codificados por el mtDNA humano en especies procariotas

En uno de los apartados de la sección de Resultados se incide en el interés de analizar también la conservación de una determinada posición de un polipéptido no sólo teniendo en cuenta proteínas ortólogas codificadas por el mtDNA de otras especies eucariotas. Al resultar la mitocondria fruto de una endosimbiosis con organismos procariotas producida hace unos 1500 millones de años, existen también proteínas ortólogas a las humanas de origen bacteriano.

Para ello, y debido a la diversidad de nombres con la que son codificadas las proteínas de organismos procariotas, fue necesaria primeramente una identificación de los nombres más representativos de estos ortólogos en las bases de datos. Esto se consiguió ejecutando un PSI-Blast (Altschul et al., 1997) desde el servidor del NCBI usando los polipéptidos humanos presentes en la base de datos RefSeq como secuencias de consulta ("query sequence") y recuperando únicamente secuencias bacterianas. Otros nombres no aparecidos tras la búsqueda con PSI-Blast pero presentes en otras publicaciones también fueron considerados.

Finalmente se recuperaron las secuencias ortólogas desde la base de datos RefSeq del NCBI. Los patrones de búsqueda en formato Entrez para cada polipéptido están descritos en la Tabla 2.

Polipéptido	Patrón de búsqueda en formato Entrez
p.MT-ND1	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND H[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoH[Title])
p.MT-ND2	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND N[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoN[Title])
p.MT-ND3	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND A[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoA[Title])
p.MT-ND4	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND M[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoM[Title])
p.MT-ND4L	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND K[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoK[Title])
p.MT-ND5	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND L[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoL[Title])
p.MT-ND6	Bacteria[Organism] AND REFSEQ AND NADH[Title] AND J[Title] AND (oxidoreductase[Title] OR dehydrogenase[Title] OR ubiquinone[Title] OR NuoJ[Title])
p.MT-CYB	Bacteria[Organism] AND REFSEQ AND ((cytochrome B[Title]) OR (apocytochrome b[Title])) NOT ((cytochrome C) OR (COX2))
p.MT-CO1	Bacteria[Organism] AND REFSEQ AND ((cytochrome C oxidase subunit I[Title]) OR (cytochrome C oxidase subunit 1[Title]))
p.MT-CO2	Bacteria[Organism] AND REFSEQ AND ((cytochrome C oxidase subunit II[Title]) OR (cytochrome C oxidase subunit 2[Title]))
p.MT-CO3	Bacteria[Organism] AND REFSEQ AND ((cytochrome C oxidase subunit III[Title]) OR (cytochrome C oxidase subunit 3[Title]))
p.MT-ATP6	Bacteria[Organism] AND REFSEQ AND ATP[Title] synthase[Title] subunit a[Title]

Tabla 2. Patrones de búsqueda en formato Entrez utilizados para la recuperación de ortólogos bacterianos desde la base de datos Refseq.

Para evitar la presencia de secuencias redundantes (como por ejemplo secuencias de la misma especie pero de diferentes cepas) se ejecutó un script en lenguaje Awk que eliminaba todas menos una. También se depuraron los resultados de la búsqueda descartando fragmentos incompletos o proteínas no relacionadas descargadas de forma incorrecta por pequeñas ambigüedades en los patrones de búsqueda definidos anteriormente. Esto se hizo estableciendo puntos de corte superiores

e inferiores en cuanto a la longitud de los polipéptidos para eliminar valores atípicos (outliers). Estos valores atípicos fueron determinados empíricamente analizando las gráficas secuenciales de longitud del conjunto de moléculas ortólogas descargadas para cada polipéptido (Tabla 3).

Polipéptido	Longitud media (aa)	Punto de corte inferior (aa)	Punto de corte superior (aa)
p.MT-ND1	382	280	484
p.MT-ND2	371	278	539
p.MT-ND3	157	100	214
p.MT-ND4	523	454	592
p.MT-ND4L	133	49	166
p.MT-ND5	650	509	791
p.MT-ND6	226	88	364
p.MT-CYTB	384	300	500
p.MT-CO1	557	404	710
p.MT-CO2	352	200	400
p.MT-CO3	275	150	400
p.MT-ATP6	325	190	560

Tabla 3. Puntos de corte superior e inferior (en número de aminoácidos, "aa") determinados empíricamente para la eliminación de secuencias con valores atípicos de longitud de cada polipéptido.

El alineamiento múltiple de secuencias ortólogas a cada polipéptido humano se ejecutó con MAFFT v.7.147b utilizando la versión más exacta "G-INS-i". El cálculo del CI para cada posición de los polipéptidos se realizó de forma similar a lo explicado anteriormente para los discriminadores del clasificador Mitoclass.1 utilizando los ficheros fasta obtenidos de los alineamientos múltiples.

5.10.2. Predicción de interacciones entre residuos

El programa MISTIC, ya detallado en el apartado relativo al discriminador 1 permite disponer de una puntuación del grado de coevolución para cada posición a través del parámetro cMI. Aunque para el cálculo numérico de dicho discriminador no es necesario estimar con qué residuos puede coevolucionar cada posición, una predicción sobre qué parejas o grupos de aminoácidos están covariando puede resultar

también de interés. Para ello hemos utilizado dos programas, PSICOV y H2r, ya que los resultados de ambos son complementarios.

PSICOV (Jones et al., 2012) tiene como objetivo la predicción de posiciones que están próximas en el espacio y que se encuentran en contacto físico directo. Las predicciones se basan en una matriz de covarianza que permite eliminar las dependencias transitivas (Figura 8). La versión descargable de PSICOV fue ejecutada con las siguientes opciones: psicov -f -p -g 0,99 -d 0,03 para permitir que el programa analizara también posiciones del alineamiento con alto contenido en gaps (99 %). Además, durante nuestros análisis utilizamos una "target precision matrix density" del 3 % en lugar de un valor fijo rho. Siguiendo las recomendaciones de los autores, fueron consideradas para un estudio posterior las primeras L/5 predicciones (siendo L la longitud del polipéptido). También fue necesario disminuir el valor MINEFSEQS definido por defecto en el programa para conseguir computar los alineamientos múltiples de p.MT-CO1 y p.MT-ND5 debido a la poca diversidad que presentaron las secuencias ortólogas.

El programa H2r (Merkl and Zwick, 2008) no elimina las dependencias transitivas al analizar las dependencias mutuas. En una primera etapa, el programa identifica parejas de residuos m,n que están interconectados con una señal fuerte de covariación. Después, para cada posición k, se realiza un recuento de sus dependencias que queda reflejado en el valor conn(k). Se ha determinado que residuos con alto valor conn(k) son frecuentemente importantes desde un punto de vista funcional o estructural (Dietrich et al., 2012). Hay que tener en cuenta que estos residuos con alto valor conn(k) no tienen por que encontrarse cerca en el espacio 3D. La versión descargable del programa fue ejecutada con los parámetros ofrecidos por defecto y los resultados fueron filtrados según recomendación de los autores: posiciones significativas serían aquellas con conn(k) > 3 y valor de bootstrap $\geq$ 0,8.

El programa H2r también se utilizó para detectar interacciones entre posiciones de polipéptidos diferentes pero presentes en el mismo complejo respiratorio. Para ello se fusionaron en uno solo todos los polipéptidos de origen genético mitocondrial pertenecientes a cada complejo como fichero de entrada para la ejecución de H2r. Esto se llevo a cabo uniendo de forma secuencial, uno a continuación del otro, los alineamientos múltiples de los polipéptidos. Así, se obtuvo un fichero con los alineamientos múltiples fusionados de p.MT-ND1, ND2, ND3, ND4, ND4L, ND5 y

ND6 (complejo I), otro con los alineamientos de p.MT-CO1, CO2 y CO3 (complejo IV) y un último fichero con los de p.MT-ATP6, ATP8 (complejo V).

Finalmente, algunas de las interacciones predichas por PSICOV con cierto interés fueron evaluadas utilizando las estructuras cristalinas de otras especies. El polipéptido humano de referencia fue alineado con la proteína homóloga utilizando el programa Cn3D tal y como ya fue descrito previamente en el apartado de caracterización de los dominios de los polipéptidos. Los residuos con posibilidad real de covariación fueron visualizados utilizando Jmol para justificar la validez de las predicciones analizando la proximidad entre los aminoácidos teniendo en cuenta una distancia de 8 Å como límite para una interacción directa.

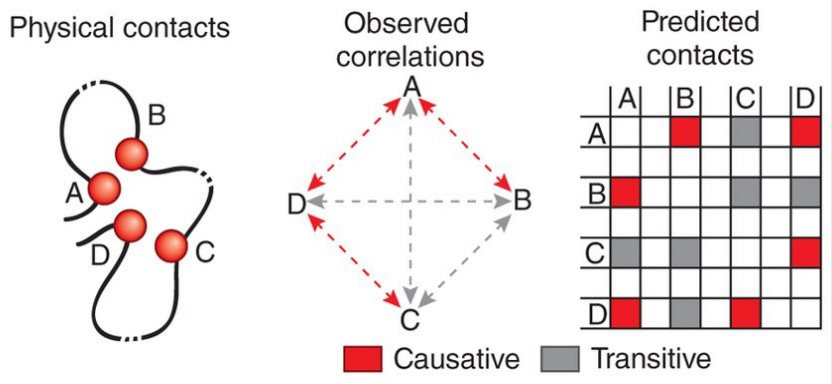

Figura 8. Diferencia entre contactos directos (causative) y contactos transitivos. Los pares de residuos A y B, A y D y D y C son interacciones directas. Por otro lado, la interacción entre A y C es transitiva debido a su interacción directa mutua con el residuo D (Marks et al., 2012).

5.10.3. Frecuencia de patogenicidad de cada tipo de aminoácido en un mismo dominio

Este parámetro es interesante para el análisis de la naturaleza de las mutaciones patológicas pero no fue incorporado como discriminador del clasificador ya que sus valores numéricos dependen del número de mutaciones pertenecientes a la clase patológica o a la clase neutra y esta información no puede ser utilizada para entrenar el predictor ya que los discriminadores deben ser independientes de la clase de mutación.

El parámetro calcula el cociente entre substituciones patológicas frente al total de substituciones (patológicas y neutras) para cada uno de los veinte aminoácidos en un mismo dominio. El número de substituciones se extrae de la base de datos de variantes conocidas mdmv.1 descrita previamente. A modo de ejemplo, para el aminoácido ácido

glutámico (E) se efectuaría un cociente entre el número de cambios que han resultado patológicos del ácido glutámico a otro aminoácido en el dominio transmembrana frente al total de cambios (patológicos y neutros) de ácido glutámico a otro aminoácido en dicho dominio. Este parámetro tiene en cuenta el número de veces que aparece cada uno de los veinte aminoácidos en el conjunto de los trece polipéptidos codificados por el mtDNA. Los valores se representan en forma de porcentaje.

5.10.4. Frecuencia de patogenicidad de un cambio particular en un mismo dominio

Al igual que el anterior discriminador, esta información no puede ser utilizada para entrenar el predictor ya que los discriminadores deben ser independientes de la clase de mutación. A pesar de ello, aporta pistas valiosas sobre la naturaleza patológica de los cambios y su dependencia del entorno bioquímico en el que se encuentra el aminoácido.

Este parámetro calcula el cociente entre el número de veces que se ha descrito un cambio particular como patológico frente al total de veces (tanto patológicas como neutras) que dicho cambio en dicho dominio aparece listado en la base de datos mdmv.1. A modo de ejemplo, el número de veces que el cambio E-K (ácido glutámico a lisina) ha sido descrito como patológico dentro del dominio transmembrana frente al número total de cambios E-K (patológicos y no patológicos) clasificados en la base de datos mdmv.1 para ese dominio. Este parámetro tiene en cuenta el número de veces que aparece el cambio en el conjunto de los trece polipéptidos codificados por el mtDNA. Los valores se representan en forma de porcentaje.

5.10.5. Calculo del grado de conservación en secuencias humanas

Para determinar si una substitución encontrada en un paciente es la responsable de la mitocondriopatía, uno de los criterios de patogenicidad clásicos define que la mutación no debe aparecer en la población control. Este criterio ha sido tenido en cuenta en la depuración de mutaciones catalogadas en MITOMAP para la generación de nuestra base de datos mdmv.1. El número de mutaciones clasificadas en MITOMAP se actualiza continuamente así como el número de genomas mitocondriales humanos completos secuenciados. A modo de ejemplo, en abril del 2014 incluía más de 27,000 secuencias. Sin embargo, este número supone solamente una pequeña fracción del total

de la población humana (unos 7000 millones de personas). En el futuro, conforme el número genomas mitocondriales secuenciados aumente, dispondremos de una población muestral más representativa del total. El número de secuencias analizadas para cada polipéptido varía desde las 21155 de p.MT-CYB a las 24437 de p.MT-ATP8. Estas diferencias se deben seguramente al hecho de que además de secuencias obtenidas de genomas completos también hemos descargado de GenBank fragmentos de secuencias de genes individuales publicadas para casos de estudios de patologías vinculadas a polipéptidos concretos.

En nuestro trabajo hemos descargado todas las secuencias disponibles de los trece polipéptidos codificados por el mtDNA humano incluidos en GenBank. Debido a la gran similaridad entre las secuencias efectuamos el alineamiento múltiple de todas las secuencias pertenecientes al mismo polipéptido a través del programa MAFFT version v7.147b con la opción "auto" (Katoh and Standley, 2013). Esta opción no es la más precisa pero el número tan elevado de secuencias hacía necesario seleccionar un método de alineamiento rápido. Además, al tratarse de un conjunto de secuencias humanas con un reducido número de polimorfismos entre ellas, la complejidad del alineamiento era pequeña y no se requería el uso de opciones más sofisticadas.

A continuación cada fichero de alineamiento generado se usó como argumento de entrada de un script escrito en Perl y Awk para calcular la frecuencia absoluta de los aminoácidos presentes en cada posición del polipéptido. También se calculó el CI , definido como la frecuencia relativa del aminoácido presente en cada posición de los polipéptidos presentes en la secuencia genómica mitocondrial humana de referencia (revised Cambridge reference sequence, rCRS, NC_012920.1).

6. Resultados y discusión

6.1. Base de datos mdmv.1

En material y métodos se describen los criterios de patogenicidad que hemos usado para el filtrado de las mutaciones descritas hasta el momento de este trabajo (año 2015) desde la web MITOMAP. La base de datos mdmv.1 consta finalmente de 2835 mutaciones, 57 de ellas patológicas (Tabla 4) (fichero anexo1.xls). Este sería el primer resultado de nuestro trabajo, puesto que una base de datos correctamente depurada de mutaciones no sinónimas de los polipéptidos codificados por el mtDNA no estaba disponible y es una de las premisas clave para el correcto desarrollo del predictor.

Predictores ampliamente utilizados utilizan la base de datos HGMD (Human Genome Mutation Database). Hemos evaluado dicha base de datos y verificado que no incluye variantes no sinónimas del mtDNA siendo en su totalidad, variantes nucleares. Otros programas han escogido la base de datos HumDiv. De nuevo, hemos verificado que incluye un número muy reducido de variantes no sinónimas del mtDNA. Además, comprobamos que en dicha base de datos aparecen catalogadas como patológicas algunas mutaciones del mtDNA que en realidad no lo son.

gen	posición AA	WT AA	M AA	dominio	fenotipo
MT-ATP6	1	M	T	intermembrana	patológico
MT-ATP6	105	A	P	transmembrana	patológico
MT-ATP6	109	W	R	transmembrana	patológico
MT-ATP6	155	A	P	transmembrana	patológico
MT-ATP6	156	L	P	transmembrana	patológico
MT-ATP6	156	L	R	transmembrana	patológico
MT-ATP6	162	A	V	transmembrana	patológico
MT-ATP6	167	G	S	transmembrana	patológico
MT-ATP6	168	H	R	transmembrana	patológico
MT-ATP6	169	L	P	transmembrana	patológico
MT-ATP6	170	L	P	transmembrana	patológico
MT-ATP6	217	L	P	transmembrana	patológico
MT-ATP6	217	L	R	transmembrana	patológico
MT-ATP6	220	L	P	transmembrana	patológico
MT-ATP6	222	L	P	matriz	patológico
MT-CO1	232	Q	K	transmembrana	patológico
MT-CO2	1	M	T	intermembrana	patológico
MT-CO2	135	L	P	intermembrana	patológico
MT-CO3	195	S	P	transmembrana	patológico
MT-CYB	34	G	S	transmembrana	patológico
MT-CYB	278	Y	C	intermembrana	patológico

MT-ND1	2	P	S	transmembrana	patológico
MT-ND1	24	E	K	transmembrana	patológico
MT-ND1	25	R	Q	transmembrana	patológico
MT-ND1	52	A	T	matriz	patológico
MT-ND1	56	F	L	matriz	patológico
MT-ND1	59	E	K	matriz	patológico
MT-ND1	110	S	N	transmembrana	patológico
MT-ND1	128	A	T	transmembrana	patológico
MT-ND1	131	G	S	transmembrana	patológico
MT-ND1	132	A	T	transmembrana	patológico
MT-ND1	143	E	K	transmembrana	patológico
MT-ND1	195	R	Q	matriz	patológico
MT-ND1	214	E	K	matriz	patológico
MT-ND1	215	Y	H	matriz	patológico
MT-ND1	289	L	M	transmembrana	patológico
MT-ND2	71	L	P	transmembrana	patológico
MT-ND3	34	S	P	matriz	patológico
MT-ND3	45	S	P	matriz	patológico
MT-ND4	158	L	P	transmembrana	patológico
MT-ND4	340	R	H	matriz	patológico
MT-ND4	340	R	S	matriz	patológico
MT-ND4L	65	V	A	transmembrana	patológico
MT-ND5	124	F	L	transmembrana	patológico
MT-ND5	236	A	T	transmembrana	patológico
MT-ND5	243	V	I	transmembrana	patológico
MT-ND5	253	V	A	transmembrana	patológico
MT-ND5	312	L	P	transmembrana	patológico
MT-ND5	393	D	N	transmembrana	patológico
MT-ND5	393	D	G	transmembrana	patológico
MT-ND6	25	P	L	transmembrana	patológico
MT-ND6	36	G	S	transmembrana	patológico
MT-ND6	60	L	S	transmembrana	patológico
MT-ND6	63	M	V	transmembrana	patológico
MT-ND6	64	M	V	transmembrana	patológico
MT-ND6	64	M	I	transmembrana	patológico
MT-ND6	72	A	V	transmembrana	patológico

Tabla 4. Substituciones de aminoácidos con fenotipo patológico presentes en la base de datos mdmv.1. WT y M simbolizan respectivamente el aminoácido salvaje (wild type) y el mutante causante de la patología. También se muestra el dominio en el que se encuentra ubicada la posición del polipéptido afectada.

6.2. Caracterización de los dominios de los trece polipéptidos codificados por el mtDNA humano

El cálculo del discriminador 3 utilizado por el clasificador Mitoclass.1 requiere conocer previamente el dominio en el que se encuentra una determinada posición del polipéptido. Siguiendo la metodología desarrollada en el apartado de material y métodos, hemos localizado las posiciones que se encuentran en cada uno de los tres

dominios (intermembrana, transmembrana y matriz) para cada uno de los trece polipéptidos (Tabla 5).

Polipéptidos Humanos	**Localización N-terminal**	**Dominios: Segmentos del polipéptido**
p.MT-ND1	TM	IM: 84-102; 159-178; 179-192; 241-256; 314-318 TM: 1-27; 74-83; 103-122; 128-158; 219-240; 257-276; 283-313 M: 28-73; 123-127; 193-218; 277-282
p.MT-ND2	IM	IM: 1-2; 44-48; 123; 170-174; 223-237; 300-313 TM: 3-21; 26-43; 49-80; 93-122; 124-144; 148-169; 175-190; 200-222; 238-273; 276-299; 314-332 M: 22-25; 81-92; 145-147; 191-199; 274-275; 333-347
p.MT-ND3	TM	IM: 79-83 TM: 1-21; 53-78; 84-107 M: 22-52; 108-115
p.MT-ND4	IM	IM: 1-7; 42-68; 112-114; 172-185; 247-252; 303-305; 388; 448-459 TM: 8-18; 23-41; 69-88; 96-111; 115-134; 142-171; 186-207; 228-246; 253-275; 282-302; 306-336; 355-387; 389-414; 428-447 M: 19-22; 89-95; 135-141; 208-227; 276-281; 337-354; 415-427
p.MT-ND4L	IM	IM: 1; 53-54 TM: 2-21; 25-52; 55-82 M: 22-24; 83-98
p.MT-ND5	IM	IM: 1-14; 64-83; 134-136; 191-192; 265-271; 320-321; 401-405; 463-494 TM: 15-33; 42-63; 84-107; 114-133; 137-156; 162-190; 193-240; 241-264; 272-291; 298-319; 322-350; 368-400; 406-431; 449-462; 495-521; 594-603 M: 34-41; 108-113; 157-161; 292-297; 351-367; 432-448; 522-593
p.MT-ND6	IM	IM: 1-2; 47-48; 111-135 TM: 3-21; 25-46; 49-74; 91-110; 136-157 M: 22-24; 75-90; 158-174
p.MT-CYB	M	IM: 53-75; 131-174; 245-286; 341-344; TM: 33-52; 76-104; 110-130; 175-201; 222-244; 287-309; 322-340; 345-376; M: 1-32; 105-109; 202-221; 310-321; 377-380
p.MT-CO1	M	IM: 42-51; 118-141; 214-229; 285-298; 359-371; 436-448 TM: 12-41; 52-83; 95-117; 142-170; 183-213; 230-262; 270-284; 299-327; 336-358; 372-396; 407-435; 449-478

		M: 1-11; 84-94; 171-182; 263-269; 328-335; 397-406; 479-513
p.MT-CO2	IM	IM: 1-14; 89-227
		TM: 15-46; 62-88
		M: 47-61
p.MT-CO3	M	IM: 36-40; 108-128; 184-191; 256-261
		TM: 16-35; 41-65; 73-107; 129-153; 156-183; 192-224; 233-255
		M: 1-15; 66-72; 154-155; 225-232
p.MT-ATP6	IM	IM: 1-7; 71-85; 178-184
		TM: 8-25; 52-70; 86-121; 151-177; 185-221
		M: 26-51; 122-150; 222-226
p.MT-ATP8	IM	IM: 1-7
		TM: 8-24
		M: 25-68

Tabla 5. Dominios de los polipéptidos codificados por el mtDNA. IM, TM y M son las abreviaturas de dominio intermembrana, transmembrana y matriz, respectivamente.

Al analizar los dominios hemos detectado que, de los 13 polipéptidos, 8 de ellos disponen su extremo N-terminal en el dominio intermembranoso. Únicamente tres polipéptidos lo presentan en el dominio matriz y en dos de los casos el extremo N amino-terminal se ubica en el propio dominio transmembrana. Esto supone que diez de los trece polipéptidos se insertan en la membrana interna mitocondrial siguiendo el sentido desde el espacio intermembrana/transmembrana hacia la matriz.

El método seguido para la caracterización de dominios depende de dos factores: la existencia de una estructura cristalina de la proteína ortóloga a la humana y la calidad del alineamiento. En algunos de los casos estos cristales proceden de bacterias (*Thermus thermophilus* para todas las subunidades del complejo I codificadas por el mtDNA y *Escherichia coli* para p.MT-ATP6). Estas especies están muy alejadas filogenéticamente de los humanos y sus secuencias podrían presentar grandes diferencias que afectaran al alineamiento y por consiguiente, a una correcta caracterización de los dominios. Por ejemplo, el homólogo bacteriano de p.MT-ND2 consta de 14 hélices transmembrana mientras que el polipéptido humano sólo conserva 11 de dichas hélices (Birrell and Hirst, 2010). El hecho de que nuestro protocolo de caracterización de dominios haya detectado 11 hélices transmembrana en la secuencia humana de p.MT-ND2 tras su alineamiento con la secuencia ortóloga bacteriana, corrobora que el método de trabajo seguido para la caracterización de dominios es adecuado.

6.3. Análisis del índice de conservación interespecífico

En los tres discriminadores utilizados por Mitoclass.1 se utilizan directa o indirectamente parámetros relacionados con la conservación de un determinado aminoácido o de una substitución concreta. Por ello, debido a su gran importancia en el desarrollo de nuestro clasificador, dedicamos este capítulo a un análisis biológico minucioso del índice de conservación interespecífico (CI) como justificación de por qué lo hemos seleccionado como discriminador de patogenicidad.

6.3.1. Selección del índice de conservación (CI) como parámetro para el estudio de conservación evolutiva

El índice de conservación (CI), es un criterio comúnmente utilizado para la determinación de patogenicidad en los polipéptidos codificados por el mtDNA (DiMauro and Schon, 2001; Montoya et al., 2009). Para desarrollar este análisis, el único requerimiento es la ejecución de un alineamiento múltiple de secuencias de polipéptidos ortólogos al humano. Actualmente, no existe un método que destaque sobre los demás para la correcta cuantificación de la conservación de un determinado residuo y han aparecido muchas propuestas en la última década sin que ninguna haya sido aceptada como método de referencia (Capra and Singh, 2007; Valdar, 2002). Entre las metodologías existentes, podemos citar la entropía de Shannon (Durbin et al., 1998), la entropía de Von Neumann (Caffrey et al., 2004), la entropía relativa (Wang and Samudrala, 2006) o el algoritmo Rate4Site (Pupko et al., 2002). Además, se ha demostrado que el resultado predictivo depende de factores relacionados con el alineamiento utilizado como el número de secuencias incluidas, el método de alineamiento o el grado de homología de las secuencias seleccionadas (Hicks et al., 2011).

Para confirmar la calidad del CI como parámetro informativo del grado de conservación del residuo, realizamos un estudio comparativo con otros métodos descritos en un trabajo previo (Capra and Singh, 2007): sum of pairs, JS divergence, relative entropy, property entropy, property relative entropy y Shannon entropy. Para este estudio utilizamos el alineamiento múltiple obtenido para el péptido p.MT-ND1 que en aquel momento contaba con 4114 secuencias ortólogas en organismos eucariotas. Los coeficientes de correlación de Spearman obtenidos entre el CI y el resto

de métodos variaron de 0,73 a 0,88, verificando de este modo la validez del CI como parámetro para cuantificar el grado de conservación.

6.3.2. Asociación entre el CI y el grado de importancia funcional/estructural de una posición

Una posición clave para el polipéptido estará altamente conservada a través de la evolución debido a que una mutación en esta posición será eliminada por selección negativa. A modo de ejemplo, las histidinas 83, 97, 182 y 196 de p.MT-CYB y las de las posiciones 61, 376 y 378 de p.MT-CO1 que unen los grupos hemo requeridos para las reacciones de transferencia de electrones en la cadena respiratoria (Kim et al., 2012) muestran valores de CI iguales o mayores a 99,77 % de acuerdo a nuestro panel de unas 5000 secuencias ortólogas (fichero anexo2.xls).

Además, al analizar el CI medio de los aminoácidos afectados por mutaciones patológicas (78,9 %) y compararlo con el de las mutaciones neutras (41,3 %), verificamos que es significativamente diferente (P = 6,657e-15). Así pues, una mutación altamente conservada es a priori interesante. A pesar de ello no se puede concluir que lo contrario no lo sea. De las 57 mutaciones que hemos determinado como patológicas en nuestra base de datos mdmv.1, observamos que aparecen 10 mutaciones (17,5 % del total de patológicas) asociadas a posiciones con un CI menor de 50 %. Un ejemplo de mutación patológica poco conservada es la mutación m.14484T>C asociada a LHON. La substitución de una metionina por una valina en la posición 64 (Brown et al., 1992; Johns et al., 1992; Mackey and Howell, 1992) presenta un CI de 12,7 % en 5177 especies ortólogas analizadas. Esto sugiere que substituciones con baja conservación no deben ser excluidas en la búsqueda de una mutación candidata para una patología. Una explicación de su patogenicidad podría ser la presencia de coevolución con otros aminoácidos del mismo o distinto polipéptido (Schmidt et al., 2001). La presencia de esta coevolución será tratada posteriormente en este trabajo. Por otro lado, encontramos 396 mutaciones catalogadas como neutras en mdmv.1 (14,2 % del total de neutras) con CI mayor de 90 %, es decir, altamente conservadas. Esto podría explicarse por el hecho de que dichas mutaciones sean levemente deletéreas y que, en humanos, por ser una especie joven desde el punto de vista evolutivo todavía no haya pasado tiempo suficiente para que la selección natural las elimine (Kryukov et al., 2007).

6.3.3. Influencia del número de secuencias ortólogas en el análisis de la conservación

El número de secuencias genómicas mitocondriales de organismos eucariotas incluidas en las bases de datos ha crecido exponencialmente y continúa haciéndolo en la actualidad. Sin embargo, esta información se utiliza de forma parcial y sesgada en el cálculo de parámetros como el CI, siendo su valor absolutamente dependiente del número de secuencias ortólogas consideradas (Figura 9).

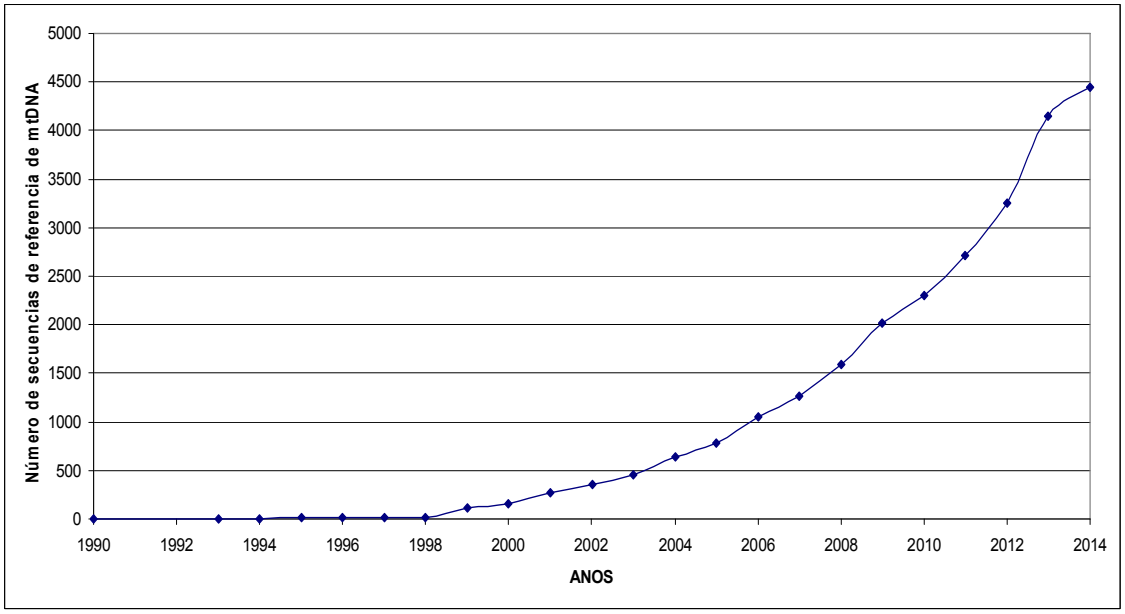

Figura 9. Frecuencia acumulada del número de secuencias de referencia de mtDNA de especies eucariotas incluidas en la base de datos RefSeq del NCBI desde 1990 hasta el año 2014.

A modo de ejemplo podemos citar la mutación m.3460G>A asociada a neuropatía óptica hereditaria de Leber (LHON), en la que una alanina es sustituida por una treonina en la posición 52 de p.MT-ND1. El aminoácido alanina apareció en 7 especies de las 9 analizadas en el trabajo citado. Esto suponía un CI de 77,7 %. Los autores consideraron por ello que dicha posición estaba conservada a través de la evolución (Huoponen et al., 1991). Sin embargo, si analizamos el total de especies eucariotas secuenciadas hasta 2015 (5165 especies para el péptido p.MT-ND1), dicho cambio presenta una conservación mucho menor, con un CI del 36,6 %.

Otro ejemplo aparece en la transversión m.15434C>A encontrada en un paciente con cardiomiopatía dilatada en la que se sustituye una leucina de la posición 230 de p.MT-CYB (Zarrouk Mahjoub et al., 2012). Esta leucina fue también considerada

altamente conservada a través de la evolución con un CI de 97,7 % según dichos autores. Para el cálculo del CI se utilizaron 43 especies de primates. Sin embargo, el CI obtenido en nuestro estudio (analizando 4988 secuencias de dicho péptido de especies diferentes, desde protistas a humanos) fue del 66 %. En el año 2013 aparecieron 8 artículos presentando nuevas mutaciones patológicas en el mtDNA humano. En todos estos trabajos se utilizaron 30 especies o menos para calcular la conservación de la posición involucrada en el cambio patológico y excepto en un caso (CI =99,3 %), en el resto el CI resultó ser de 100 %. Es interesante comentar que ninguna de estas mutaciones obtuvo un CI de 100 % en nuestro panel de aproximadamente 5000 especies y que únicamente tres de ellas poseían un CI superior a 90 %. Además, una de ellas mostró un CI inferior a 20 %.

Otra táctica utilizada por diversos programas y que puede afectar a la estimación de la conservación de una posición es la utilización de herramientas para eliminar la redundancia del conjunto de homólogos obtenido. El objetivo consiste en seleccionar a un conjunto de secuencias representativas y eliminar aquellas que son altamente parecidas entre sí. Para ello, el usuario define un grado de similitud como punto de corte, de forma que el programa agrupa las secuencias y elimina aquellas con alta identidad. Un programa ampliamente utilizado para ello es CD-HIT (Li et al., 2001). Así, los parámetros definidos por el usuario para CD-HIT pueden afectar al cálculo de la conservación ya que el número de secuencias finales y su grado de homología dependerán del agrupamiento realizado.

6.3.4. Dependencia del método de recuperación de secuencias de las bases de datos

Un amplio número de predictores utilizan BLAST (Basic Local Alignment Search Tool) como método para la descarga de secuencias homólogas a una dada desde las bases de datos. Sin embargo, esto supone que secuencias parálogas (originadas por un evento de duplicación dentro del mismo genoma) o pseudogenes nucleares de los polipéptidos codificados por el mtDNA (NUMTs) (Tsuji et al., 2012) podrían ser recuperados por su alta identidad con la secuencia humana y utilizados en el posterior alineamiento provocando un sesgo en el cálculo de la conservación. Además, ortólogos con poca homología respecto a la secuencia humana correspondientes a especies distantes evolutivamente podrían no ser descargados por BLAST y se perderían para el posterior análisis de la conservación.

La incidencia de estos sesgos sobre el valor numérico final de la conservación dependerá de los parámetros que elijamos para la ejecución de BLAST: la base de datos, el número máximo de secuencias, el punto de corte para el valor esperado (E-value), la matriz de alineamiento utilizada, el coste de introducción de gaps, etc... Para evitar este sesgo debido al número de secuencias ortólogas elegido para el análisis de la conservación, nuestro estudio ha incluido todos los polipéptidos ortólogos al humano codificados por el mtDNA procedentes de los organismos presentes en la base de datos RefSeq. El número de secuencias descargadas se encuentra entre las 4668 para el polipéptido p.MT-ATP8 y las 5177 de p.MT-ND6. Estas diferencias en el número de secuencias descargadas entre polipéptidos son debidas por un lado a razones biológicas ya que los 13 polipéptidos del mtDNA humano no se encuentran codificados en los genomas mitocondriales de todos los organismos debido a que algunos de ellos se codifican a través de genes nucleares (Wallace, 2007). Además, algunos organismos pueden presentar más de una copia del mismo gen en su mtDNA o codificar para más de una proteína dentro del mismo gen (Tabla 6).

Código RefSeq del Genoma completo mitocondrial	**Organismo**	**Genes con más de una copia**
NC_006354	*Todarodes pacificus*	*MT-COX1, COX2, COX3, ATP6, ATP8*
NC_007893	*Watasenia scintillans*	*MT-COX1, COX2, COX3, ATP6, ATP8*
NC_009093	*Metaseiulus occidentalis*	*MT-ND1, ND4, ND4L, ND5, CYTB, COX1, COX2, COX3, ATP6, ATP8*
NC_009493	*Fusarium graminearum*	*MT-ND1, ND2, ND3, ND4L, ND5, CYTB, COX1, COX2, COX3, ATP6*
NC_017855	*Daucus carota subsp. sativus*	*MT-ND3, ND5, CYTB, COX3*
NC_016423	*Bathyteuthis abyssicola*	*MT-COX1, COX2, ATP8*
NC_014338	*Proteromonas lacertae*	*MT-ND1, ND2, ND3, ND4L, ND6*
NC_010636	*Sthenoteuthis oualaniensis*	*MT-COX1,COX2, COX3,ATP6, ATP8*

Tabla 6. Algunos ejemplos de secuencias genómicas de la base de datos RefSeq de especies con mtDNA completo secuenciado en las que aparece más de una copia del mismo gen.

La diferencia observada en el número de ortólogos descargados para cada gen también puede ser explicada por razones bioinformáticas ya que existe un número minoritario de secuencias que no fueron incluidas en la base de datos RefSeq por los autores con el mismo nombre que el resto de sus ortólogos y que por ello no pueden ser

descargadas con nuestro protocolo automático de descarga de secuencias. Esto se debe a que nuestro código en lenguaje Perl recupera todas aquellas secuencias que contengan un mismo nombre (campo "protein name" en los ficheros GenBank).

El problema de las ambigüedades a la hora de establecer el nombre de genes y proteínas en las bases de datos está ampliamente documentado y requerirá que estas sean poco a poco depuradas en el futuro (http://www.uniprot.org/docs/nameprot.txt). Un ejemplo de este efecto podemos encontrarlo al analizar el número de secuencias descargadas para los péptidos p.MT-CYB y p.MT-CO1. Ambos están codificados en todos los genomas mitocondriales eucariotas (Wallace, 2007) y sin embargo, aparecen en 4988 y 4829 especies respectivamente. Además de que el número no es el mismo, este es inferior a las 5177 secuencias descargadas para p.MT-ND6. Se conocen diferentes formas de denominar a p.MT-CO1. En nuestro programa de descarga de secuencias hemos utilizado el nombre de la proteína "cytochrome c oxidase subunit I". Sin embargo, a modo de ejemplo, para el nombre alternativo "Cytochrome C Oxidase I" han aparecido dos secuencias que no son descargadas utilizando nuestro método. Esto mismo podría ocurrir para los otros doce genes. Ante la gran cantidad de nombres alternativos y su escasa representación numérica, hemos preferido concentrar la recuperación de secuencias empleando su denominación más común.

Otro problema que debería resolverse en el futuro es el sesgo presente en las bases de datos debido a que algunas categorías taxonómicas se encuentran sobrerrepresentadas mientras que otras contienen actualmente pocas especies secuenciadas. Esto puede ser debido a varios motivos como por ejemplo que en la comunidad científica exista más interés actualmente por ciertos organismos. En un análisis del número de organismos llevado a cabo en 2012, observamos que mientras el filo de los cordados incluía mas de 100 secuencias de cada clase (mamíferos, aves, reptiles, anfibios, etc...), otros reinos como Fungi, Plantae o Protista aportaban menos de 100 especies cada uno. Por otro lado, dentro de los cordados, el reino Animalia contenía más de 2700 secuencias (90 %) sobre un total de algo más de 3000 especies eucariotas presentes por aquel entonces en la base de datos RefSeq.

6.3.5. Uso del CI en el control de calidad de genomas mitocondriales humanos

El CI también puede utilizarse como control de calidad del proceso de secuenciación del mtDNA. El conocimiento de la filogenia del mtDNA humano permite

descubrir potenciales errores producidos durante el proceso de secuenciación si estos afectan a mutaciones que definen haplogrupos (Bandelt et al., 2005). Estos haplogrupos trazan la ascendencia matrilineal hasta los orígenes de la especie humana en África (Eva mitocondrial, hace 200000 años) y desde allí, a su subsiguiente dispersión por toda la superficie del planeta (Soares et al., 2009).

Sin embargo, posibles errores en mutaciones privadas en las puntas del árbol filogenético del mtDNA son más difíciles de detectar. Las mutaciones privadas son mutaciones raras encontradas en una única familia o en una población muy reducida. Estas mutaciones suelen transmitirse a unos cuantos miembros familiares pero no aparecen en generaciones futuras. Para estos casos, el CI puede aportar indicios sobre si una nueva mutación es susceptible de aparecer en un individuo sano, considerando que posiciones muy conservadas a lo largo de la evolución serían peores candidatas de albergar mutaciones privadas.

Por ejemplo, la transición m.15290C>T se ha descrito en un individuo sano, MA161(Kumar et al., 2011). Esta transición provoca la sustitución de una histidina en la posición 182 por una tirosina en el polipéptido p.MT-CYB. Esta histidina es importante en la unión del grupo hemo que participa en las reacciones de transferencia de electrones desarrollada por dicho polipéptido. Además, está conservada con un CI de 99,8 % en 4988 especies distintas por lo que dicho cambio sería interesante confirmarlo a través de una nueva secuenciación de dicha posición.

Por otro lado, la transición m.15002G>A ha sido descrita como una mutación presente en dos ramas internas del árbol filogenético del mtDNA (Gonder et al., 2007). Esta transición provoca la sustitución de una glicina en la posición 86 por una serina del polipéptido p.MT-CYB. Esta glicina está conservada con un CI de 97,0 % en 4988 especies distintas. Recientemente se ha publicado que el conjunto de secuencias utilizadas en dicho trabajo es de baja calidad (Yao et al., 2008) y sería interesante verificar esas secuencias para confirmar dicha transición.

6.3.6. Análisis del CI en especies procariotas

Es importante remarcar que en la base de datos RefSeq contenida en GenBank, existían en 2014 más de 3500 secuencias ortólogas de animales pero menos de 500 considerando únicamente las de hongos, plantas y protistas juntas. Este sesgo en el tipo de organismos elegidos para computar el CI podría afectar a la evaluación de la

importancia funcional de una sustitución particular. Por ello, decidimos estudiar este efecto incluyendo un análisis de secuencias ortólogas filogenéticamente distantes de la humana. En concreto, decidimos utilizar secuencias ortólogas procedentes de organismos procariotas.

El número de secuencias recuperadas para cada polipéptido presentó mucha variabilidad, desde 0 (p.MT-ATP8) hasta 2353 (p.MT-ATP6) (fichero anexo3.xls). El polipéptido p.MT-ATP8 es una subunidad supernumeraria encontrada únicamente en ATP sintasas de origen animal y hongos. Una subunidad supernumeraria es un polipéptido adicional que no se encuentra en las ATP sintasas bacterianas. Estas enzimas bacterianas parecen contener la mínima cantidad posible de subunidades requeridas para la actividad de la ATP sintasa (Hong and Pedersen, 2004). Las grandes diferencias encontradas en el número de secuencias de los polipéptidos contenidos en el complejo respiratorio I (384 para p.MT-ND4L y 631 para p.MT-ND5) son más difíciles de explicar porque aparentemente todos los complejos I bacterianos conocidos contienen las siete subunidades (desde p.MT-ND1 hasta p.MT-ND6) (Friedrich and Böttcher, 2004). Algo similar ocurre con las subunidades del complejo IV con 786 y 1200 secuencias de p.MT-CO3 y p.MT-CO1 respectivamente. Para este complejo IV también parece que todas las bacterias contienen las tres subunidades (p.MT-CO1, CO2 y CO3). La razón de estas diferencias es, probablemente, bioinformática. Nuestro criterio de recuperación de secuencias desde las bases de datos, menos depurado que para el caso de organismos eucariotas, no consigue la descarga completa de todas las secuencias posibles debido a la gran diversidad de nombres con que estos péptidos aparecen en GenBank.

Analizando la conservación en bacterias se podrían descubrir posiciones muy conservadas en eucariotas que sin embargo, no lo son tanto en procariotas. Para verificar esto, hemos escogido el polipéptido p.MT-DN1 por ser el gen del que más mutaciones patológicas hay confirmadas en la base de datos mdmv.1. Para dicho polipéptido hemos estudiado la conservación de todas aquellas variantes de la base de datos mdmv.1 relativas a posiciones con CI mayor o igual a 95 % en eucariotas (Tabla 7). Se puede observar que las mutaciones patológicas en dicho gen que cumplen ese requisito (7 substituciones) presentan una conservación muy similar tanto en eucariotas (99,00 %) como en procariotas (95,86 %). Sin embargo, para las mutaciones neutras por encima de 95 % de CI en eucariotas (19 sustituciones) se aprecia una disminución de la conservación en procariotas (84,33 %) (Tabla 7).

Para el polipéptido p.MT-ATP8 ya hemos comentado que no existen ortólogos de origen bacteriano. De los polipéptidos p.MT-ND2, ND3, ND4L, ND6 y CO3 no existen actualmente mutaciones patológicas publicadas con CI mayor de 95 %. Para el resto de polipéptidos: p.MT-ND4, ND5, CYB, CO1, CO2 y ATP6 hemos repetido el análisis descrito para p.MT-ND1. Los resultados globales confirman esta mayor bajada de conservación en procariotas para el grupo de las mutaciones neutras sobre el de patológicas. El cálculo total se ha realizado con una media ponderada teniendo en cuenta el número de mutaciones patológicas/neutras de cada gen. En los tres genes de los que existen un mayor número descrito de mutaciones patológicas confirmadas (p.MT-ND1, ND5, ATP6) se observa claramente esta tendencia. Sin embargo, en los otros genes, con sólo una o dos mutaciones patológicas clasificadas (p.MT-ND4, CYB, CO1 y CO2) aparece el comportamiento contrario, con una subida del CI en los organismos procariotas para las mutaciones neutras. Haría falta disponer de un número mayor de mutaciones patológicas para poder verificar esta tendencia y confirmar la hipótesis de que posiciones muy conservadas en eucariotas tienen menos probabilidad de albergar mutaciones patológicas en caso de observarse una bajada de conservación en organismos bacterianos.

	p.MT-ND1	p.MT-ND4	p.MT-ND5	p.MT-CYB	p.MT-CO1	p.MT-CO2	p.MT-ATP6	total
Nº mutaciones patológicas	7	2	6	2	1	1	6	25
CI mutaciones patológicas en eucariotas	99	99,2	98,4	99,6	98,1	97,1	98,2	98,6
CI mutaciones patológicas en bacterias	95,8	69,4	89,8	24	73	2,2	75,6	77
Nº mutaciones neutras	19	13	21	37	56	15	15	176
CI mutaciones neutras en eucariotas	98	97,7	98,1	97,4	98,1	98,5	97,7	97,9
CI mutaciones neutras en bacterias	82,7	80,7	84,8	19,6	47,3	37	62	52,6
Número de secuencias bacterianas	393	624	631	856	1200	1160	2353	

Tabla 7. Valores de índice de conservación (CI) para aquellas posiciones con CI en eucariotas superior a 95 % para los polipéptidos codificados por el mtDNA humano presentes en organismos procariotas.

6.3.7. Análisis del CI medio de los polipéptidos

La mutación LHON m.14495A>G provoca una substitución de leucina por serina en la posición 60 de p.MT-ND6 (Chinnery et al., 2001). El CI de esta leucina es de 77,8 % analizado en 5177 especies. No es por tanto una posición altamente conservada aunque si analizamos el CI medio de cada uno de los trece polipéptidos codificados por el mtDNA, éste oscila entre 21,2 % y 78,5 % (Figura 10). Así pues, el CI de dicha mutación es similar al CI medio del polipéptido más conservado (p.MT-CO1 con 78,5 %) pero mucho mayor que el CI medio de p.MT-ND6 (27,9 %). Tal vez relativizar la conservación de una posición frente a la de su polipéptido resultara ser mejor discriminador que el CI aislado. Para hacernos una idea de la relación entre el CI de las mutaciones patológicas y de los polipéptidos que las albergan, hemos analizado los dos genes con mayor número de cambios patológicos (p.MT-ND1 y p.MT-ATP6), con 15 sustituciones cada uno. En el caso de p.MT-ND1, el CI medio del polipéptido es

de 59 % y el CI medio de las mutaciones patológicas asociadas a dicho gen es del 78 %.

En el caso de p.MT-APT6 el CI medio del polipéptido es del 48 % y el CI medio de sus mutaciones patológicas es del 90 %. Parece que existe una correlación inversa entre el CI medio del polipéptido y el CI de sus mutaciones patológicas. Es decir, en polipéptidos menos conservados serían potencialmente patológicas únicamente las posiciones muy conservadas. Sin embargo, para corroborar esto sería necesario un número mayor de mutaciones patológicas publicadas. El escaso número presente en el actualidad (57) no permite realizar un estudio estadísticamente adecuado y por ello hemos descartado utilizar algún parámetro derivado de esta observación como posible discriminador.

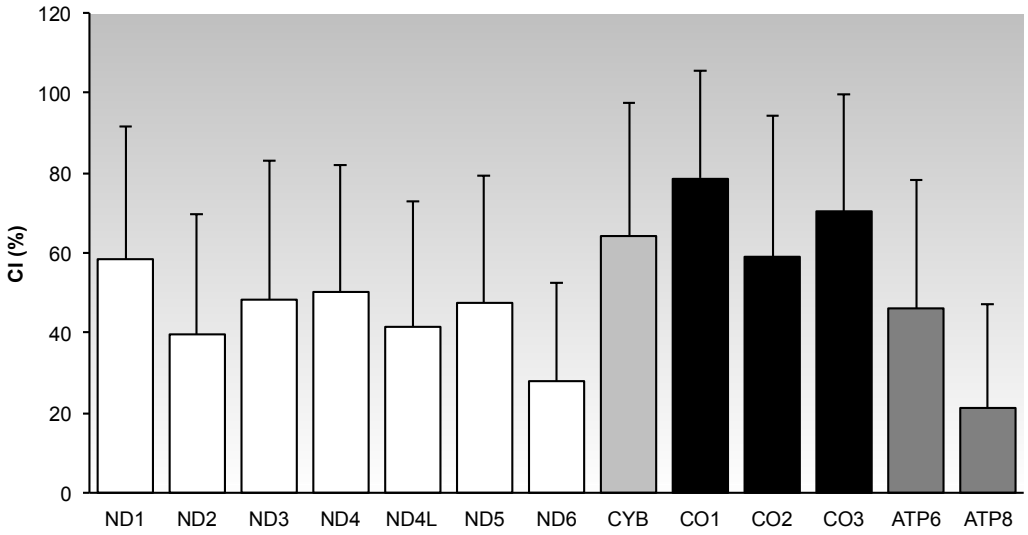

Figura 10. Conservación media de los trece polipéptidos codificados por el mtDNA humano medida según el índice de conservación (CI).

6.3.8. Análisis de la conservación por dominios dentro de un mismo polipéptido

Existe mucha variación en la conservación entre polipéptidos tal y como hemos verificado en el apartado anterior. Esto también podría ocurrir para los dominios de un mismo polipéptido. Para corroborar estas diferencias se analizó el CI de cada segmento del polipéptido p.MT-ND1(Figura 11).

Puede apreciarse que el CI medio de los bucles del espacio intermembrana es del 30,6 % mientras que en la matriz aumenta hasta el 74,6 %. Esta elevada conservación del dominio matriz coincide con el hecho de que bastantes mutaciones patológicas de

dicho polipéptido (6 de 15) se localizan en bucles extramembranosos del lado de la matriz (Valentino et al., 2004). Recientemente se ha propuesto además que el dominio matriz de esta subunidad podría ser importante en la interacción de los polipéptidos del complejo I pertenecientes a la membrana interna y la matriz y que contribuiría a la actividad del sitio de unión a coenzima Q (Efremov and Sazanov, 2012).

Hemos analizado si estas diferencias de conservación observadas en los dominios de p.MT-ND1 se muestran de manera global en el total de los trece polipéptidos pero no las hemos encontrado (Tabla 8). El CI medio de cada dominio es muy similar (55, 58 y 51 % para el dominio intermembrana, transmembrana y matriz, respectivamente). Tampoco existen grandes variaciones al analizar el CI medio de todas las mutaciones neutras de la base de datos mdmv.1 por dominios. A pesar de que las diferencias no son importantes sí que se observa que el dominio matriz es el menos conservado mientras que el dominio transmembrana es el que más conservado está (valor p = 7,8e-07).

Curiosamente hemos obtenido los valores mayores de CI para las mutaciones patológicas del dominio intermembrana (95 %). Además, este valor de CI es muy diferente y mayor que el obtenido para los otros dos dominios. Este resultado deberá confirmarse en el futuro cuando dispongamos de un número mayor de mutaciones patológicas publicadas ya que actualmente sólo existen cuatro mutaciones descritas en el dominio intermembrana y de ellas, dos afectan al aminoácido inicial del polipéptido, siendo por ello una posición a priori muy conservada.

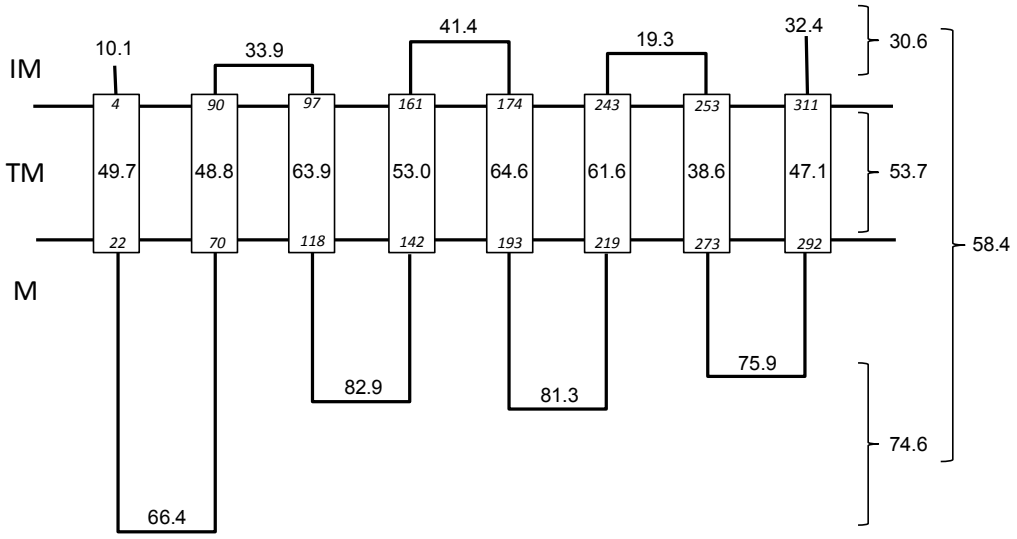

Figura 11. CI por dominios de p.MT-ND1. IM, TM y M son las abreviaturas de dominio intermembrana, transmembrana y matriz, respectivamente. En cursiva se indican las posiciones numéricas de los extremos de cada segmento transmembrana.

Dominio	CI medio de mutaciones patológicas en el dominio	CI medio de todas las posiciones del dominio	CI medio de mutaciones neutras en el dominio
IM	95	55	41
TM	78	58	43
M	73	51	35

Tabla 8. índice de conservación (CI) por dominios teniendo en cuenta los trece polipéptidos codificados por el mtDNA humano medido según el parámetro CI. Las mutaciones neutras y patológicas se han extraído de la base de datos mdmv.1. IM, TM y M son las abreviaturas de dominio intermembrana, transmembrana y matriz, respectivamente.

6.4. Estudio de la naturaleza de las mutaciones patológicas presentes en mdmv.1

De forma previa a la selección de los parámetros discriminadores más adecuados para el clasificador Mitoclass.1 efectuamos un estudio pormenorizado de las mutaciones clasificadas en la base de datos mdmv.1. En concreto, analizamos la naturaleza de los aminoácidos salvajes a los que afectaban las mutaciones patológicas, así como la naturaleza de los tipos de substitución de carácter deletéreo. Todo ello teniendo también en cuenta el dominio en el que se encontraban las mutaciones.

6.4.1. Análisis de la frecuencia de patogenicidad de cada tipo de aminoácido dentro del mismo dominio

Con este parámetro tratamos de establecer un valor numérico que mida la importancia funcional de cada aminoácido particular dentro de cada uno de los tres dominios establecidos, siempre teniendo en cuenta la información contenida en la base de datos mdmv.1. El hecho de que cada uno de los tres dominios en los que se encuentran los polipéptidos presente propiedades físico-químicas distintas puede hacer que la substitución del aminoácido sea más susceptible de resultar patológica en uno de ellos. Para analizar correctamente el parámetro, hemos determinado previamente la frecuencia absoluta/relativa de cada aminoácido en cada dominio (Figura 12), así como el número de mutaciones patológicas y no patológicas que le afectan (Tabla 9).

Por un lado, al estudiar la frecuencia relativa de cada tipo de aminoácido por dominios observamos por ejemplo que:

1) La prolina (P) es muy poco frecuente en el dominio transmembrana (TM), por el hecho de que dicho aminoácido afecta a la estructura de las alfa hélices.
2) Otros aminoácidos poco frecuentes en el dominio TM son arginina (R), ácido aspártico (D), glutamina (Q) y ácido glutámico (E). Todos ellos son polares y por ello más infrecuentes en un entorno hidrofóbico como el dominio transmembrana.
3) La arginina (R) y la lisina (K) se encuentran con mayor frecuencia en el dominio matriz. Ambos son polares y están cargados positivamente, coincidiendo con el hecho de que los residuos con carga positiva suelen encontrarse en el interior de la matriz (von Heijne, 1992).
4) La alanina (A) y la valina (V) están menos representados en el dominio matriz que en los otros dos dominios. Ambos son hidrofóbicos y de pequeño tamaño.

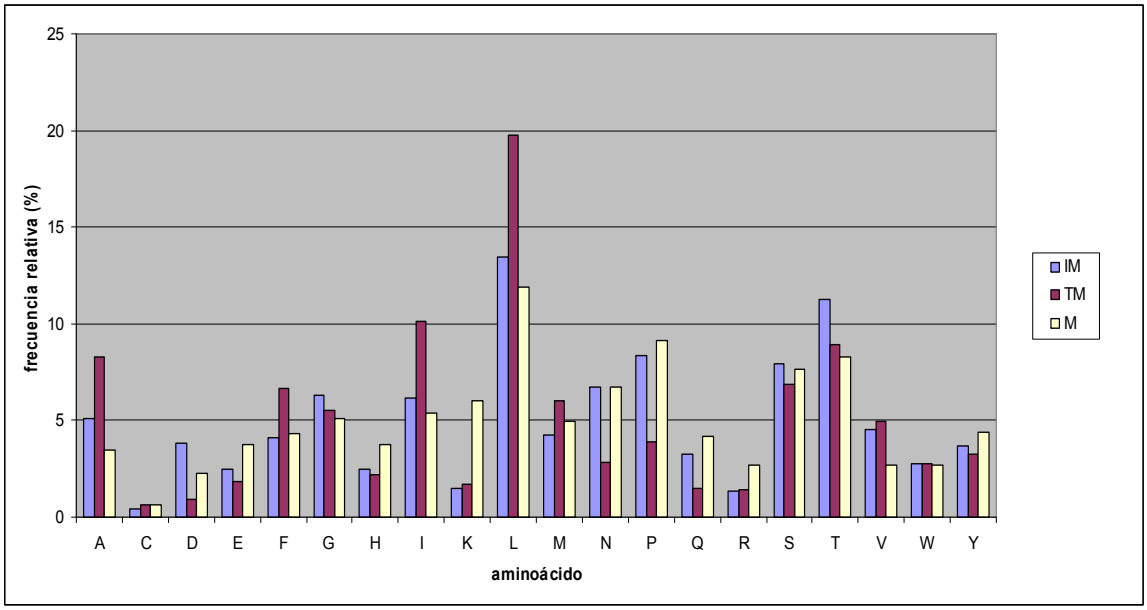

Figura 12. Frecuencia relativa de cada aminoácido en cada uno de los tres dominios. IM, TM y M son las abreviaturas de dominio intermembrana, transmembrana y matriz respectivamente. Los aminoácidos están representados por su código de una letra.

Analizando los valores de la frecuencia de patogenicidad obtenidos observamos varios aspectos importantes:

1) Las mutaciones patológicas de la base de datos mdmv.1 no están distribuidas de forma homogénea por los tres dominios con una mayor tendencia a ubicarse en el dominio transmembrana.
2) Los aminoácidos afectados por mutaciones patológicas tampoco son los mismos en cada uno de los dominios. Por ejemplo, en el caso del dominio intermembrana las cuatro mutaciones patológicas descritas afectan a 3 tipos de residuos: leucina, metionina y tirosina (L, M, Y). En el dominio matriz las 11 mutaciones patológicas afectan a 7 residuos distintos (A, E, F, L, R, S, Y). Finalmente, en el dominio transmembrana, el que muestra una mayor frecuencia de substituciones patológicas, aparecen afectados hasta 14 aminoácidos distintos (A, D, E, F, G, H, L, M, P, Q, R, S, V, W). (Tabla 9).
3) Las mutaciones patológicas que afectan a leucina se encuentran en los tres dominios, probablemente debido a que es el tipo de aminoácido que aparece con mayor frecuencia en todos ellos.
4) Existen cinco tipos de aminoácidos para los cuales no se han descrito todavía cambios patológicos. Se trata de C, I, K, N y T. En el caso del residuo cisteína

(C), seguramente no se han encontrado mutaciones patológicas debido a que es el aminoácido menos frecuente en todos los dominios. Curiosamente, el aminoácido treonina (T) se encuentra entre los tres más frecuentes de cada dominio y es uno de los tres aminoácidos para los que existen más mutaciones descritas, aunque ninguna de ellas ha sido catalogada como patológica. Algo parecido ocurre con la isoleucina (I). La isoleucina es el segundo aminoácido más frecuente del dominio transmembrana y también uno de los que ha experimentado más mutaciones neutras. La asparagina (N) es también uno de los cinco aminoácidos más frecuentes en los dominios hidrofílicos (intermembrana y matriz) y el segundo en número de mutaciones neutras en el dominio matriz. Estos resultados sugieren que el papel de estos aminoácidos (I, N y T) no es muy relevante en la funcionalidad o estructura de los polipéptidos. El residuo lisina (K) apenas está presente en los dominios intermembrana y transmembrana, explicando por ello que no tenga asociadas mutaciones patológicas. En cambio, en el dominio matriz es uno de los de mayor frecuencia (el sexto) pero apenas aparece afectado por mutaciones neutras.

5) La arginina (R) es uno de los aminoácidos menos frecuentes del dominio matriz y sin embargo muestra mutaciones patológicas, así como el valor más alto para esta frecuencia de patogenicidad, remarcando así su importancia funcional.
6) Los cuatro aminoácidos con valores más altos para esta frecuencia de patogenicidad dentro del dominio hidrofóbico transmembrana son aminoácidos hidrofílicos como el ácido aspártico (D), glutamina (Q), ácido glutámico (E) y arginina (R). Como el dominio transmembrana de los polipéptidos codificados por el mtDNA participa en la translocación de protones hacia el espacio intermembrana y de vuelta al dominio matriz, estos aminoácidos podrían ser importantes en este proceso.

	IM					TM					M				
AA	**N**	**%**	**M**	**P**	**F**	**N**	**%**	**M**	**P**	**F**	**N**	**%**	**M**	**P**	**F**
A	37	5,1	37	0	0	192	8,3	177	7	4,0	26	3,5	27	1	3,7
C	3	0,4	1	0	0	14	0,6	8	0	0	5	0,7	4	0	0
D	28	3,8	25	0	0	21	0,9	9	2	22,2	17	2,3	16	0	0
E	18	2,5	7	0	0	42	1,8	23	2	8,7	28	3,8	15	2	13,3
F	30	4,1	22	0	0	154	6,7	104	1	1,0	32	4,3	18	1	5,6
G	46	6,3	25	0	0	128	5,5	60	4	6,7	38	5,1	15	0	0
H	18	2,5	12	0	0	51	2,2	22	1	4,5	28	3,8	26	0	0
I	45	6,2	44	0	0	234	10,1	318	0	0	40	5,4	55	0	0
K	11	1,5	1	0	0	39	1,7	9	0	0	45	6,0	17	0	0
L	98	13,4	44	1	2,3	457	19,7	205	12	5,8	89	11,9	46	1	2,2
M	31	4,3	20	2	10,0	140	6,1	134	3	2,2	37	5,0	36	0	0
N	49	6,7	49	0	0	65	2,8	59	0	0	50	6,7	70	0	0
P	61	8,4	31	0	0	90	3,9	41	2	4,9	68	9,1	41	0	0
Q	24	3,3	8	0	0	35	1,5	11	1	9,1	31	4,2	20	0	0
R	10	1,4	2	0	0	33	1,4	14	1	7,1	20	2,7	13	3	23,1
S	58	8,0	55	0	0	159	6,9	117	2	1,7	57	7,6	61	2	3,3
T	82	11,2	73	0	0	207	8,9	201	0	0	62	8,3	79	0	0
V	33	4,5	32	0	0	114	4,9	147	3	2,0	20	2,7	18	0	0
W	20	2,7	1	0	0	64	2,8	19	1	5,3	20	2,7	7	0	0
Y	27	3,7	17	1	5,9	75	3,2	36	0	0	33	4,4	31	1	3,2
	729		506	4		2314		1714	42		746		615	11	

Tabla 9. Frecuencia de patogenicidad de cada aminoácido en un mismo dominio. IM, TM y M son las abreviaturas de dominio intermembrana, transmembrana y matriz. AA, N, %, M, P y F indican los aminoácidos, la frecuencia absoluta, la frecuencia relativa, el número de mutaciones, el número de mutaciones patológicas y la frecuencia de patogenicidad, respectivamente.

A pesar de que este parámetro no puede ser utilizado como discriminador para el entrenamiento de un clasificador debido a que está basado en un ratio entre mutaciones de clase patológica y clase neutra (ya comentado en material y métodos), su análisis nos ha ofrecido información muy interesante.

6.4.2. Análisis de la frecuencia de patogenicidad de un cambio particular dentro del mismo dominio

Este discriminador trata de cuantificar la importancia funcional de una substitución específica dentro del mismo dominio polipeptídico. El hecho de que una misma sustitución presente diferente efecto en función del dominio ha sido anteriormente documentado (Tourasse and Li, 2000) y muchas propiedades fisicoquímicas pueden resultar importantes a la hora de determinar el fenotipo (O. et al., 2013; Saha et al., 2012) (Figura 13).

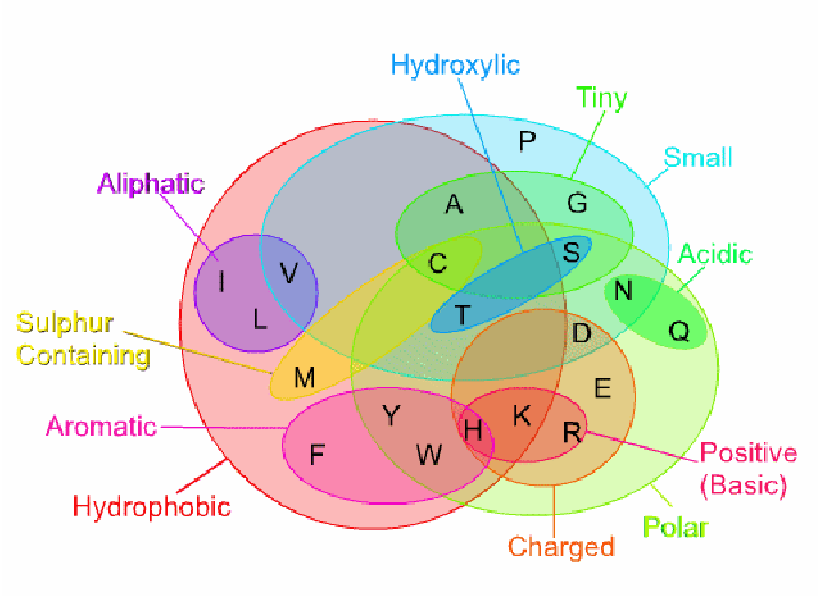

Figura 13. Propiedades con las que habitualmente se clasifica a los aminoácidos (O. et al., 2013).

Al revisar todas las sustituciones presentes en la base de datos mdmv.1 hemos observado que:

1) Los tres dominios muestran cambios de leucina a prolina (L-P) asociados a fenotipos patológicos. La leucina, y también la alanina, favorecen la generación de alfa-hélices pero la prolina provoca la pérdida de puentes de hidrógeno en los péptidos, desestabiliza la estructura secundaria e introduce un pliegue en la alfa hélice (Frank S Cordes, 2002; Vanhoof et al., 1995). Todos los cambios L-P patológicos (diez) menos uno afectan a la estructura secundaria de alfa hélices. De esos diez, ocho se encuentran en el dominio transmembrana. La presencia o no de las otras dos substituciones en el interior de alfa-hélices se verificó estudiando estructuras ortólogas cristalinas y alineando dichas proteínas con las secuencias humanas. Uno de ellos, el cambio L222P en p.MT-ATP6 se encuentra en la frontera entre la matriz y el dominio transmembrana y ocupa la parte final de una alfa-hélice. El otro cambio, L135P aparecido en p.MT-CO2 se encuentra en el espacio intermembrana y es el único de los diez en el que la variante se sitúa en una posición sin estructura secundaria bien definida.
2) En el dominio transmembrana, los cinco cambios con valores mayores para este discriminador (≥ 20 %) son: D-G, D-N, L-R, Q-K y R-Q. Todos los cambios patológicos con valores más altos para este discriminador en el dominio transmembrana implican, curiosamente, perdida (D-G, D-N y R-Q) o ganancia de carga eléctrica (L-R, Q-K) de tres aminoácidos cargados (D, R, K), que

además, se encuentran entre los cinco menos frecuentes de este dominio (Tabla 10) por su marcado carácter hidrofóbico. El dominio transmembrana de estos polipéptidos participa en el movimiento de protones a través de la membrana interna de la mitocondria. Por ello, esta clase de sustituciones afectando a la carga eléctrica de esas posiciones podría influir en el correcto proceso de translocación de protones.

3) El dominio transmembrana es el único en el que aparecen dos variantes en las que es sustituido el aminoácido prolina (P). La prolina es el aminoácido con menor predisposición para formar alfa hélices (Pace and Scholtz, 1998). Sin embargo, está documentado que a veces este aminoácido aparece en el principio de una alfa hélice tal y como ocurre en la posición 2 de p.MT-ND1, posición para la que hay descrita una substitución patológica de prolina a serina (P2S). En cambio, para la posición 25 de p.MT-ND6, vinculada a un cambio patológico (P25L), el aminoácido parece ocupar una zona más intermedia en la alfa hélice, tal vez generando un giro necesario que la mutación destruye.
4) El aminoácido glicina (G) es el segundo con menor propensión helicoidal tras la prolina y es el dominio transmembrana el único en el que aparece un cambio patológico vinculado a este aminoácido y además, cuatro veces.
5) Analizando los aminoácidos con mayor propensión helicoidal, encontramos al grupo denominado "MALEK" (metionina, alanina, leucina, ácido glutámico y lisina). Curiosamente, si revisamos el conjunto de variantes patológicas que aparecen en el dominio transmembrana pero no en el dominio matriz, encontramos numerosos casos de aminoácidos de este grupo: A-V, M-V, M-I, L-S, A-P, L-R. Esto quedaría justificado por su importancia para la formación de alfa hélices.
6) Algo parecido a lo comentado para el dominio transmembrana ocurre en el dominio matriz, en el que los cuatro cambios con valores más altos para este discriminador (≥ 20 %) producen pérdida de carga eléctrica (R-H, R-Q, R-S) o cambio de carga (E-K) de tres aminoácidos cargados (R, E, K), dos de ellos (R y E) entre los menos frecuentes de este dominio. Algunos aminoácidos cargados establecen interacciones eléctricas con aminoácidos de polipéptidos periféricos y son importantes en el ensamblaje de los complejos OXPHOS. Así, dos sustituciones E-K en p.MT-ND1 (p.E59K y p.E214K) se ha demostrado que

afectan al correcto ensamblaje del complejo respiratorio I (Kirby et al., 2004; Malfatti et al., 2007).

7) En el dominio matriz aparecen dos cambios de serina (S) a prolina (P) que, sospechosamente, se producen en dos posiciones muy cercanas del mismo polipéptido. Se trata de p.MT-ND3 y el cambio S-P aparece en las posiciones 34 y 45 siendo ambos patológicos. No es posible saber la estructura secundaria de esa zona del polipéptido porque no alinea bien con el ortólogo del que se conoce su estructura cristalina pero analizando la secuencia del polipéptido, la posición 36 está ocupada por una prolina al igual que las posiciones 43 y 46. Este peculiar aminoácido se encuentra habitualmente asociado a giros en los que la proteína debe cambiar de dirección. Por ello, tal vez el cambio de serina a prolina introduzca nuevos giros inesperados en la molécula en una región en la que ya existen prolinas en zonas muy próximas.
8) Curiosamente, existen tres variantes patológicas: Y-H, R-H y R-S que sólo aparecen en el dominio matriz. La frecuencia de aparición de Y y R tanto en el dominio matriz como en el transmembrana es muy parecida y muy baja por lo que habría que averiguar porque su importancia es mayor en la matriz. Además, no existe hasta la fecha ninguna mutación patológica vinculada al dominio transmembrana en el que el cambio sea a histidina, mientras que en el matriz sí.
9) En el caso del dominio intermembrana no aparecen cambios con valor para este discriminador ≥ 20 %. Las tres substituciones patológicas descritas son: L-P, que parece no afectar a una alfa hélice (comentado previamente), M-T y Y-C. Además, ninguno de los tres cambios confirmados hasta la fecha como patológicos en este dominio contiene aminoácidos cargados eléctricamente tal y como ocurría en el resto de dominios. Por otro lado, el cambio con valor mayor para el discriminador en el espacio intermembrana (M-T, 18,2 %) afecta directamente al aminoácido inicial en los dos polipéptidos afectados (p.MT-ATP6 y p.MT-CO2). Aunque se mantenga la terminología típica de una variante, en realidad el polipéptido podría no llegar a sintetizarse al verse afectado el triplete de iniciación por lo que el aminoácido metionina no sería substituido realmente por treonina. Esto ha sido confirmado para el caso del gen p.MT-CO2 aunque todavía no está claro para la variante del gen p.MT-ATP6 (Clark et al., 1999; Ware et al., 2009).

IM					TM					M				
WT AA	**M AA**	**P**	**T**	**F**	**WT AA**	**M AA**	**P**	**T**	**F**	**WT AA**	**M AA**	**P**	**T**	**F**
M	T	2	11	18,2	Q	K	1	2	50,0	R	S	1	1	100
Y	C	1	6	16,7	D	G	1	3	33,3	R	H	1	3	33,3
L	P	1	9	11,1	D	N	1	4	25,0	E	K	2	8	25,0
					L	R	2	9	22,2	R	Q	1	4	25,0
					R	Q	1	5	20,0	S	P	2	16	12,5
					L	P	8	41	19,5	F	L	1	10	10,0
					G	S	4	24	16,7	L	P	1	13	7,7
					W	R	1	7	14,3	Y	H	1	13	7,7
					E	K	2	15	13,3	A	T	1	14	7,1
					A	P	2	17	11,8					
					H	R	1	9	11,1					
					S	N	1	11	9,1					
					L	S	1	12	8,3					
					P	L	1	13	7,6					
					P	S	1	16	6,3					
					M	I	1	19	5,3					
					M	V	2	43	4,7					
					V	A	2	43	4,7					
					A	V	2	48	4,2					
					S	P	1	27	3,7					
					A	T	3	88	3,4					
					L	M	1	39	2,6					
					V	I	1	43	2,3					
					F	L	1	53	1,9					

Tabla 10. Frecuencia de patogenicidad de un cambio particular dentro del mismo dominio. IM, TM y M son las abreviaturas de dominios intermembrana, transmembrana y matriz, respectivamente. WT AA, M AA, P, T y F indican el tipo de aminoácido salvaje, el aminoácido mutante, el número de sustituciones patológicas, la frecuencia de cada sustitución particular y el valor numérico de la frecuencia de patogenicidad, respectivamente. No se han descrito mutaciones patológicas para otras sustituciones diferentes (fichero anexo1.xls). Los colores rojo y azul indican aminoácidos con carga negativa y positiva respectivamente. El color verde indica una sustitución de un aminoácido que favorece la estructura de alfa hélice frente a otro (Prolina, P) que introduce un pliegue en la hélice.

Al igual que para el parámetro comentado previamente en el punto anterior, esta frecuencia tampoco puede ser usada para el diseño de un clasificador. Sin embargo, aporta información valiosa que es necesario analizar.

6.5. Análisis de los discriminadores escogidos para el clasificador Mitoclass.1

A continuación, justificamos las razones biológicas y bioinformáticas que nos han hecho seleccionar los tres parámetros discriminadores en los que se basa el clasificador Mitoclass.1.

6.5.1. Discriminador 1. CI + cMI en Eucariotas

Un valor alto de conservación en una posición ofrece una pista sobre su importancia funcional. Sin embargo, una posición poco conservada no significa que no sea importante. Así, es posible que una sustitución (A-B) en una posición X de un polipéptido esté compensada a través de la evolución por un cambio (C-D) en otra posición distinta Y del mismo o de otro polipéptido (Sandler et al., 2014). Este hecho podría permitir que un nuevo aminoácido (B) quedara fijado en una posición X ofreciendo así dicha posición una conservación baja para A, aun siendo importante funcionalmente. Si esta consideración es cierta, algunas mutaciones patológicas con CI bajo mostrarán valores altos de coevolución. En nuestro estudio hemos utilizado una variante normalizada del parámetro cMI (cumulative mutual information) generado por el programa MISTIC como representativo de la coevolución. Así pues, el uso combinado del CI y el cMI (en nuestro caso la suma de ambos) permite disponer de forma numérica de un discriminador que presenta valores altos para aquellos casos en los que la mutación afecta a una posición muy conservada o, en caso de no estarlo, representa una probabilidad grande de coevolución con otros residuos.

Para justificar el uso del cMI en el discriminador 1 del clasificador Mitoclass.1 hemos realizado una comparación de los valores de cMI normalizados obtenidos por las poblaciones de mutaciones neutras y patológicas para distintos rangos de conservación (según el CI) (Figura 14). Puede observarse que un gran número de mutaciones patológicas presentan conservaciones elevadas según lo esperable. Para este grupo de mutaciones patológicas el grado de coevolución es bajo porque es poco probable que alberguen cambios polimórficos. La gráfica presenta valores de coevolución superiores para las mutaciones patológicas cuando el CI baja de 80 %. De hecho, las 19 mutaciones patológicas con CI $\leq$ 80 % ofrecen un cMI normalizado medio de 53 %. Este resultado es significativamente más alto (P = 0,0019) que el de las 2226

mutaciones no patológicas analizadas de la base de datos mdmv.1 (cMI normalizado medio = 43,8 %).

Igualmente, se observa que para CI inferiores a 60 % todas las mutaciones patológicas presentes en mdmv.1 (9 variantes) muestran valores de cMI normalizado superiores a 40 para este discriminador mientras que el conjunto de las mutaciones neutras muestra una dispersión grande por encima y debajo de dicho valor. Por ejemplo, la mutación patológica confirmada p.V65A en MT-ND4L se encuentra muy poco conservada a través de la evolución (CI= 24,7 %) pero presenta un cMI normalizado alto (78,2 %).

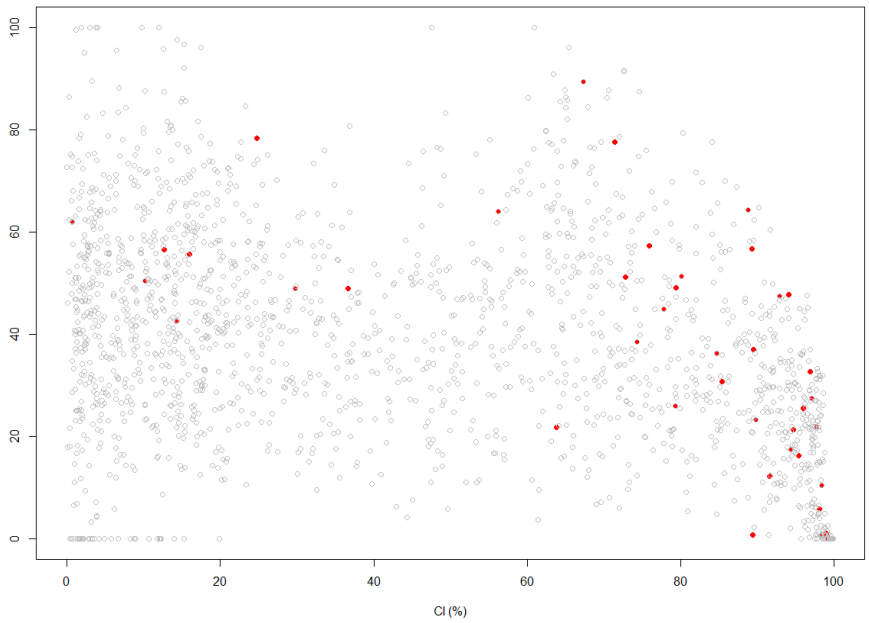

Figura 14. Comparación de los valores de cMI obtenidos por las poblaciones de mutaciones neutras (círculos grises) y patológicas (círculos rojos) para distintos rangos de conservación (según índice de conservación o CI). El eje vertical representa el cMI normalizado.

6.5.2. Discriminador 2. Conservación de los aminoácidos mutantes en cada posición de los polipéptidos

Las propiedades fisicoquímicas de los aminoácidos mutados son parámetros comúnmente analizados cuando se evalúa la patogenicidad de una mutación no sinónima. Así, suele hablarse de cambios conservativos o no conservativos cuando las propiedades del nuevo residuo son o no similares a las del aminoácido reemplazado. En cualquier caso, es difícil establecer si el cambio ha sido o no conservativo, puesto que los aminoácidos pueden clasificarse de acuerdo a muchas propiedades diferentes. A

modo de ejemplo, la reconocida base de datos AAindex database clasifica a cada aminoácido de acuerdo a 544 propiedades (Kawashima et al., 2008). Por esta razón no hemos considerado a estas propiedades por separado como discriminadores.

Teóricamente, cualquier aminoácido podría ocupar una posición sin importancia funcional/estructural de un polipéptido. Sin embargo, si esta posición es clave, sólo uno o un número muy limitado de aminoácidos químicamente relacionados podría ocupar dicha posición. Así, si la frecuencia de aparición del nuevo aminoácido es baja, más alta será la probabilidad de que el cambio resulte patológico. Esto es una consecuencia directa de la evolución por selección negativa. Al analizar el CI medio de los aminoácidos que aparecen tras una substitución en el caso de mutaciones patológicas (1,4 %), ha resultado significativamente inferior al que muestran las mutaciones neutras (8,1 %) con P=1,89e-14 en el total de mutaciones de la base de datos mdmv.1 (Figura 15). Sin embargo, hemos encontrado varias mutaciones patológicas presentes en mdmv.1 que presentan extrañamente un valor alto para este parámetro (16 % o 35 % por ejemplo). Igualmente, aparece un número muy elevado de mutaciones neutras en mdmv.1 (casi un millar) con un valor menor a 1 %. Estas frecuencias poco esperadas podrían deberse a hechos como la presencia de mutaciones compensatorias en otras partes del genoma, ya comentado anteriormente (fichero anexo4.xls).

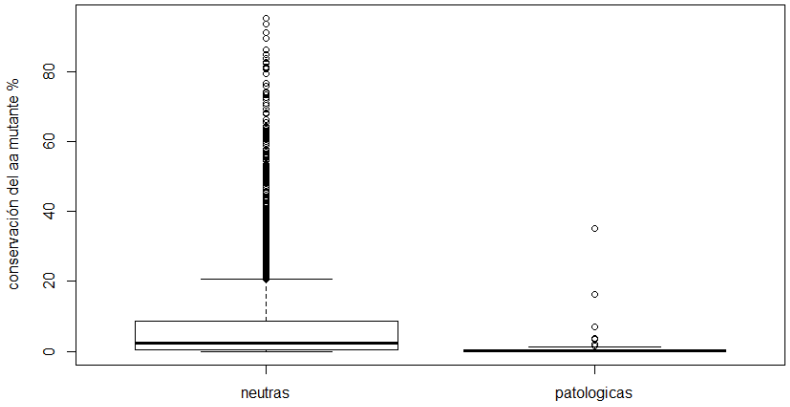

Figura 15. Boxplot representando la conservación del aminoácido mutante por posición para las dos clases de mutaciones (patológicas y neutras).

6.5.3. Discriminador 3. Frecuencia relativa de aparición de aminoácidos mutantes en un mismo dominio

Con la información presente en los alineamientos múltiples de cada polipéptido humano con el conjunto de proteínas ortólogas hemos generado las tablas de substitución por dominio tal y como se describe en material y métodos. En estas tablas se observa la conservación de cada uno de los veinte aminoácidos (diagonales), así como la frecuencia con que cada tipo de aminoácido cambia a cada uno de los otros (Tabla 11).

a) Dominio intermembrana

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	GAPS
A	50.4	0.4	1.5	0.6	2.1	1.7	0.5	5.0	0.2	4.0	3.6	1.0	4.3	0.7	0.0	9.4	5.8	4.0	0.3	0.3	4.1
C	1.3	67.4	0.2	0.3	7.7	0.7	0.7	0.3	0.5	1.8	1.7	3.7	0.1	0.2	0.1	5.6	3.1	0.5	0.6	1.9	1.5
D	0.4	0.1	73.2	7.5	0.6	0.7	1.1	0.7	0.7	0.7	0.6	6.2	0.2	1.7	0.1	1.8	1.7	0.2	0.0	0.9	0.8
E	1.1	0.1	5.4	57.5	0.4	2.1	0.3	1.4	1.9	3.5	1.7	3.6	3.1	2.2	0.2	2.3	1.9	1.0	0.1	0.5	9.6
F	0.4	0.3	2.1	1.1	60.7	0.3	0.3	2.4	0.2	8.9	1.9	0.5	2.7	0.6	0.1	1.8	0.9	1.2	3.2	5.1	5.1
G	1.3	0.2	1.9	2.4	0.4	73.9	0.1	0.3	0.6	0.6	0.4	2.2	0.2	0.3	0.1	5.0	0.9	0.6	2.1	0.4	6.2
H	0.8	2.8	0.4	1.8	1.7	0.9	64.6	0.8	0.7	1.6	1.7	6.5	0.8	1.8	0.5	2.8	1.7	0.7	0.1	3.2	4.0
I	1.1	0.3	0.4	0.8	4.2	0.8	1.0	44.4	0.2	13.6	8.4	1.4	0.7	0.1	0.1	2.1	3.4	13.0	0.2	1.0	2.9
K	0.5	0.2	0.9	3.2	0.7	0.7	2.2	1.7	58.2	2.8	1.8	3.4	1.0	7.0	0.7	3.8	3.5	1.6	0.2	1.1	4.9
L	1.6	0.3	0.5	0.3	8.7	0.8	0.3	10.1	0.4	50.1	6.2	1.4	1.8	0.9	0.0	2.9	3.8	5.2	0.4	1.5	2.9
M	1.2	0.2	0.4	0.3	2.2	0.8	1.8	4.9	1.3	10.0	52.5	4.0	0.4	0.3	0.3	3.6	6.0	4.6	0.1	0.9	4.2
N	2.0	0.2	8.1	1.8	1.5	1.5	1.9	1.1	2.6	1.8	1.3	48.4	0.9	0.8	0.2	8.5	4.8	2.2	0.5	1.9	8.1
P	3.8	0.2	0.9	0.7	0.6	1.0	0.4	1.1	0.6	1.5	0.5	1.7	70.9	1.1	0.0	5.6	3.6	1.4	0.1	0.5	3.6
Q	1.5	0.5	2.3	7.8	1.4	1.0	1.2	1.1	2.0	2.2	2.8	3.8	1.3	55.9	1.0	2.9	1.6	1.1	0.7	1.1	6.7
R	0.4	0.1	0.1	0.1	0.2	0.7	1.2	0.9	0.7	1.0	0.8	0.6	0.2	0.5	86.6	1.1	2.8	0.3	0.1	0.3	1.4
S	10.3	0.8	1.3	2.3	2.7	3.1	1.1	2.8	1.9	5.0	2.4	5.7	1.1	1.0	0.1	39.6	8.7	1.8	0.2	1.5	6.6
T	7.2	1.4	0.8	0.6	2.8	1.2	0.5	5.3	1.3	7.4	5.3	3.6	2.4	1.1	0.1	13.7	36.3	2.4	0.7	1.3	4.6
V	2.8	0.4	0.1	0.4	1.4	0.6	0.2	11.2	0.3	6.2	3.4	0.5	0.2	0.1	0.1	2.8	3.4	49.7	0.1	0.4	15.7
W	0.3	0.2	0.0	0.1	1.7	0.3	0.0	0.7	0.1	1.3	0.4	0.3	0.1	0.1	0.3	0.8	0.2	0.5	89.0	1.5	2.0
Y	0.6	0.1	0.1	0.2	13.4	0.4	2.3	1.8	0.2	4.2	0.8	1.1	0.4	0.4	0.0	1.2	1.9	1.8	0.3	67.1	1.8

b) Dominio transmembrana

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	GAPS
A	57.4	0.7	0.0	0.1	2.4	5.3	0.1	4.4	0.2	6.0	2.5	0.5	0.9	0.1	0.0	8.8	5.3	3.7	0.3	0.4	0.8
C	2.4	59.2	1.1	0.0	1.8	1.7	0.9	2.4	0.5	4.2	1.1	2.4	0.6	0.3	0.2	10.1	2.4	1.9	0.3	1.4	5.1
D	0.4	0.2	69.8	7.2	0.4	3.1	0.4	0.7	1.2	0.6	0.3	3.9	0.0	0.4	0.1	3.4	1.1	0.4	0.1	1.4	4.7
E	1.1	0.1	1.7	78.4	0.7	2.1	0.7	0.6	1.2	1.2	0.5	1.3	0.4	2.8	0.2	1.6	0.6	0.8	0.1	0.6	3.5
F	1.0	0.3	0.0	0.0	69.5	0.5	0.7	3.7	0.1	9.4	2.9	0.3	0.2	0.1	0.0	1.2	1.7	2.1	0.7	4.6	0.9
G	6.3	0.5	0.1	0.1	0.8	80.8	0.0	0.8	0.1	1.3	0.8	0.4	0.3	0.1	0.0	4.3	1.0	1.0	0.1	0.2	0.9
H	1.1	0.1	0.8	1.2	0.9	0.4	74.0	0.6	0.6	2.4	0.5	1.9	0.3	5.7	0.2	2.7	1.0	0.5	0.1	3.7	1.2
I	2.7	0.4	0.1	0.2	4.3	0.5	0.1	46.0	0.1	17.8	4.7	0.2	0.7	0.1	0.0	1.9	4.9	12.6	0.4	1.1	1.2
K	0.5	0.1	0.4	0.7	1.2	2.1	0.9	1.3	66.2	1.8	0.8	3.2	1.3	3.1	2.7	2.5	1.8	0.8	0.1	0.7	7.7
L	2.3	0.5	0.1	0.1	6.1	0.8	0.1	9.4	0.2	60.4	6.2	0.5	0.6	0.3	0.1	2.0	3.2	4.4	0.5	1.0	1.3
M	3.1	0.5	0.0	0.1	5.6	1.0	0.1	10.1	0.8	21.9	38.9	0.5	0.5	0.3	0.1	2.2	5.5	5.3	1.1	0.9	1.3
N	3.2	0.4	1.5	0.7	1.7	2.9	2.4	1.7	3.2	3.4	2.5	47.7	0.7	2.6	0.2	9.6	4.2	1.1	0.5	2.4	7.3
P	2.7	0.3	0.5	0.2	3.1	1.0	0.4	2.5	0.5	4.0	1.7	0.9	65.1	0.3	0.1	4.8	3.5	1.4	0.3	1.1	5.3
Q	0.8	0.2	0.7	2.7	0.9	0.9	2.2	1.1	2.3	5.7	1.8	1.1	0.9	63.2	5.9	3.1	1.5	1.0	0.3	1.0	2.6
R	0.5	0.2	0.1	0.3	0.8	1.8	2.0	0.9	2.4	0.9	0.3	2.0	0.1	1.2	80.5	1.3	0.6	0.6	0.2	0.6	2.7
S	10.0	1.2	0.2	0.3	2.5	5.3	0.3	2.7	0.9	4.1	2.3	2.3	0.7	0.9	0.2	53.1	6.7	2.2	0.5	0.8	2.7
T	8.5	1.5	0.1	0.2	2.9	1.7	0.3	7.2	0.8	10.1	5.0	1.3	0.9	0.3	0.2	9.4	40.6	5.5	0.3	0.9	2.2
V	3.2	0.4	0.0	0.0	3.4	1.9	0.0	14.8	0.1	9.8	5.2	0.4	0.8	0.1	0.0	2.1	3.7	50.7	0.4	0.9	1.9
W	0.6	0.2	0.1	0.2	3.8	1.5	0.1	1.8	0.1	3.0	0.9	0.2	0.2	0.1	0.1	1.2	0.5	1.3	79.9	2.5	1.8
Y	0.6	1.0	0.4	0.6	8.9	0.3	3.1	1.6	0.6	3.6	1.1	2.3	0.2	0.7	0.1	2.2	1.5	1.1	2.5	64.9	2.6

c) Dominio matriz

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	GAPS
A	44.6	0.4	0.1	0.1	1.2	10.8	0.2	4.0	1.8	2.6	1.7	1.2	1.8	1.0	0.2	12.3	7.6	2.7	0.1	1.0	4.6
C	0.5	64.5	0.0	0.1	1.4	1.3	1.1	0.3	0.1	2.1	0.2	0.5	0.2	0.1	0.1	2.7	2.9	0.2	0.7	4.2	16.7
D	0.2	0.0	72.4	10.9	0.1	0.2	3.4	0.7	0.5	0.3	0.1	5.2	0.3	0.5	0.5	1.0	2.7	0.3	0.0	0.1	0.6
E	0.6	0.1	2.0	71.0	0.7	0.6	0.4	1.0	1.8	3.7	0.7	3.4	2.0	2.1	0.5	1.7	1.8	2.7	0.1	0.5	2.6
F	1.9	0.2	0.1	0.1	58.3	0.4	0.6	4.8	0.1	15.2	2.4	0.6	0.1	0.2	0.0	1.0	1.3	1.6	0.3	8.3	2.4
G	2.0	0.2	0.2	0.3	0.9	77.0	0.1	0.5	0.7	0.9	0.8	2.9	2.0	0.2	0.5	2.8	2.1	1.0	1.3	0.2	3.4
H	0.9	0.1	3.4	0.6	1.4	1.3	55.9	0.8	3.3	2.5	0.6	5.9	1.4	4.6	0.3	3.9	2.2	0.4	0.3	2.2	7.9
I	4.2	0.6	0.1	0.2	3.4	0.7	0.3	35.5	1.5	17.7	4.5	1.2	1.7	0.6	0.3	3.6	7.4	11.3	0.2	1.8	3.5
K	1.1	0.2	2.5	4.1	1.0	2.7	1.1	1.3	45.3	2.2	1.3	7.2	3.4	4.9	4.4	4.7	3.9	1.0	0.2	1.2	6.2
L	1.8	0.3	0.1	0.4	5.4	1.7	0.8	7.9	1.5	49.0	5.6	2.1	1.0	1.2	0.4	3.9	4.5	4.1	1.0	1.7	5.4
M	3.0	0.2	0.1	0.3	3.0	0.4	0.5	13.3	1.8	15.2	34.1	1.0	0.6	1.0	0.7	3.9	7.0	6.9	1.1	0.9	4.9
N	2.8	0.3	3.0	1.6	1.5	3.8	5.3	1.6	5.9	2.3	1.1	44.0	1.5	2.0	2.2	9.0	4.5	1.0	0.1	2.5	4.0
P	2.7	0.1	0.5	0.9	0.9	0.8	0.4	1.3	0.9	2.3	0.9	1.9	65.2	1.2	0.2	5.2	3.4	1.1	1.5	0.6	8.0
Q	0.8	0.2	0.6	3.9	1.9	1.5	2.0	1.1	3.0	3.6	1.6	2.8	3.1	54.1	6.2	3.3	1.9	1.0	0.6	1.1	5.8
R	0.2	0.3	0.2	0.3	0.8	1.6	0.6	0.5	1.6	0.9	0.4	3.4	0.1	0.9	82.6	1.2	0.3	0.5	0.2	2.1	1.4
S	6.4	0.4	1.1	1.3	1.4	5.1	1.0	1.9	2.2	4.6	2.1	5.5	2.6	0.9	0.4	42.2	9.4	1.5	0.2	1.5	8.4
T	6.8	0.3	0.8	0.7	3.2	2.4	1.4	6.0	1.6	7.9	4.3	4.0	3.3	1.1	0.7	13.1	33.4	3.8	0.3	1.6	3.4
V	4.9	0.2	0.1	0.2	2.4	1.6	0.1	12.3	2.3	15.2	2.3	0.7	1.8	0.2	0.4	3.5	2.9	42.0	0.3	2.0	4.6
W	0.3	0.1	0.0	0.1	3.5	0.3	1.2	1.2	0.4	3.2	0.9	0.5	0.2	0.3	3.0	0.7	1.0	0.7	71.5	2.7	8.4
Y	0.6	0.2	2.8	0.9	18.4	0.3	7.5	2.1	2.2	4.3	1.8	1.9	0.5	1.5	2.9	1.6	1.1	1.4	0.2	43.2	4.5

Tabla 11. Frecuencias relativas por tipo de residuo para cada dominio. Los valores numéricos son porcentuales. La columna del título representa los aminoácidos salvajes mientras que la fila del título representa a los aminoácidos mutantes a los que puede cambiar.

Ya hemos comentado que las propiedades de cada aminoácido (carga, hidrofobicidad, tamaño, etc...) pueden determinar su importancia funcional o estructural en cada dominio. Esta importancia podría quedar reflejada en valores altos de conservación para un determinado aminoácido dentro de dicho dominio. A modo de ejemplo, el ácido glutámico (E) muestra una conservación en el dominio transmembrana del 78 %, siendo este valor, superior al que aparece en los otros dominios (57 % y 71 % en los dominios intermembrana y matriz respectivamente). El ácido glutámico es un aminoácido hidrofílico con carga eléctrica negativa y polar, muy infrecuente en dicho dominio (1,8 % de las posiciones totales del dominio, Tabla 9). Así, la presencia de este aminoácido en un dominio hidrofóbico como el transmembrana y su alta conservación podrían constatar su importancia. De hecho, existen dos mutaciones patológicas vinculadas a dicho aminoácido en el dominio transmembrana (p.E24K y p.E143K en p.MT-ND1). El aminoácido triptófano (W) también está muy conservado en el dominio intermembrana (79,9 %) y es muy infrecuente en dicho dominio (2,7 % del total, Tabla 9) pero en este caso todavía no se han descrito mutaciones patológicas vinculadas a dicho aminoácido.

En el caso de la histidina (H), hablamos de un aminoácido polar y cargado positivamente con mayor conservación en el dominio transmembrana que en el resto de dominios siendo además un aminoácido muy infrecuente en él (2,2 % de las posiciones totales, Tabla 9). Actualmente, existe una mutación clasificada como patológica en dicho dominio: pH168R en p.MT-ATP6.

También hemos observado que la arginina (R) parece ser un aminoácido importante en los tres dominios, como así demuestra su alta conservación, superior a 80 % de frecuencia en los tres. Este residuo está muy poco representado (1,3 %, 1,4 % y 2,6 % en los dominios intermembrana, transmembrana y matriz, respectivamente, Tabla 9) y ha sido asociado a varias mutaciones patológicas: p.R240H en p.MT-ND4 y p.R195Q en p.MT-ND1 ambas en el dominio matriz y p.R25Q en p.MT-ND1 en el dominio transmembrana.

Por otro lado, analizamos la frecuencia evolutiva de aparición de cada substitución en los tres dominios con la intención de verificar si una misma variante era menos frecuente en un dominio concreto, encontrando pistas así sobre su potencial interés. Estas frecuencias constituyen el parámetro discriminador 3 de los tres que constituyen el predictor Mitoclass.1. El valor numérico del discriminador se ha calculado eliminando la conservación del aminoácido salvaje y la frecuencia de aparición de los gaps de alineamiento. De este modo, el valor numérico refleja el porcentaje de aparición de cada variante con respecto a las 19 variantes posibles para cada aminoácido salvaje (Tabla 12).

a) Dominio intermembrana

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		0.97	3.30	1.27	4.68	3.81	1.09	10.99	0.48	8.74	7.92	2.27	9.42	1.45	0.04	20.61	12.81	8.71	0.72	0.71
C	4.06		0.73	0.96	24.75	2.32	2.40	0.83	1.59	5.73	5.55	11.88	0.39	0.78	0.42	18.13	9.79	1.46	1.98	6.25
D	1.71	0.40		29.01	2.27	2.70	4.15	2.56	2.52	2.56	2.44	23.86	0.96	6.49	0.24	7.04	6.69	0.85	0.17	3.37
E	3.30	0.24	16.49		1.25	6.40	1.02	4.31	5.75	10.63	5.26	10.93	9.35	6.74	0.55	7.11	5.83	3.15	0.18	1.50
F	1.29	0.91	6.27	3.17		0.82	0.95	7.09	0.73	25.97	5.54	1.57	8.03	1.72	0.18	5.41	2.71	3.46	9.27	14.92
G	6.78	0.85	9.42	11.92	2.09		0.74	1.31	3.05	3.01	2.21	11.01	0.79	1.42	0.51	25.09	4.50	2.79	10.57	1.93
H	2.60	9.00	1.37	5.78	5.56	2.79		2.53	2.27	5.07	5.34	20.85	2.65	5.71	1.48	9.04	5.30	2.20	0.40	10.04
I	2.00	0.61	0.75	1.58	8.01	1.44	1.81		0.38	25.85	15.89	2.63	1.27	0.21	0.16	3.95	6.53	24.66	0.32	1.95
K	1.45	0.41	2.33	8.71	1.97	1.88	5.94	4.53		7.63	4.78	9.19	2.70	18.92	1.77	10.39	9.52	4.46	0.44	3.00
L	3.44	0.55	1.14	0.61	18.59	1.69	0.55	21.58	0.80		13.20	3.03	3.79	1.84	0.09	6.11	7.99	11.00	0.88	3.13
M	2.70	0.55	0.84	0.74	4.99	1.74	4.15	11.35	3.10	23.03		9.28	0.98	0.63	0.66	8.32	13.90	10.69	0.22	2.15
N	4.49	0.56	18.58	4.09	3.39	3.44	4.27	2.63	5.96	4.20	2.92		2.16	1.85	0.37	19.51	11.06	4.99	1.14	4.40
P	15.09	0.70	3.57	2.58	2.53	4.01	1.63	4.39	2.32	6.04	1.97	6.73		4.47	0.18	21.86	14.00	5.64	0.34	1.93
Q	3.90	1.44	6.04	20.99	3.65	2.66	3.25	2.97	5.45	5.81	7.59	10.20	3.49		2.55	7.80	4.33	2.90	1.94	3.03
R	3.31	0.83	0.43	1.18	1.44	5.46	10.14	7.38	5.48	8.46	6.65	4.83	1.30	4.00		9.54	23.76	2.41	0.75	2.66
S	19.22	1.55	2.42	4.28	5.04	5.69	1.96	5.20	3.45	9.25	4.37	10.65	2.01	1.90	0.21		16.21	3.36	0.39	2.86
T	12.15	2.40	1.38	1.08	4.75	2.02	0.84	8.99	2.16	12.44	9.02	6.13	4.02	1.89	0.21	23.21		3.99	1.17	2.15
V	8.23	1.10	0.30	1.23	4.03	1.78	0.62	32.31	0.89	17.78	9.82	1.44	0.61	0.23	0.17	8.23	9.82		0.21	1.19
W	3.34	1.90	0.38	1.32	18.69	3.43	0.38	8.32	1.59	14.58	4.05	3.35	1.63	0.74	3.11	8.59	2.24	5.91		16.45
Y	1.77	0.39	0.38	0.80	43.13	1.31	7.47	5.62	0.71	13.41	2.49	3.50	1.39	1.18	0.09	3.72	5.99	5.70	0.95	

b) Dominio transmembrana

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		1.79	0.12	0.30	5.65	12.66	0.20	10.61	0.49	14.30	6.05	1.14	2.10	0.24	0.09	21.06	12.59	8.93	0.64	1.06
C	6.66		3.17	0.13	5.02	4.75	2.66	6.74	1.36	11.68	3.00	6.68	1.64	0.95	0.50	28.28	6.83	5.19	0.87	3.91
D	1.70	0.67		28.08	1.74	12.28	1.69	2.67	4.59	2.47	1.09	15.45	0.18	1.67	0.57	13.48	4.13	1.74	0.34	5.45
E	5.98	0.60	9.50		3.65	11.63	3.65	3.11	6.87	6.38	2.79	6.98	2.47	15.31	0.96	8.57	3.39	4.41	0.52	3.24
F	3.23	1.08	0.07	0.10		1.67	2.30	12.62	0.36	31.88	9.79	0.99	0.80	0.18	0.07	4.22	5.84	6.98	2.30	15.52
G	34.56	2.66	0.37	0.33	4.62		0.21	4.24	0.47	6.92	4.50	2.34	1.82	0.42	0.15	23.65	5.64	5.43	0.36	1.31
H	4.39	0.41	3.28	4.78	3.77	1.51		2.49	2.60	9.75	2.07	7.87	1.34	23.10	0.66	10.87	3.95	2.06	0.31	14.79
I	5.09	0.84	0.23	0.31	8.18	1.00	0.11		0.20	33.81	8.86	0.47	1.28	0.20	0.04	3.63	9.22	23.86	0.68	1.99
K	2.00	0.46	1.35	2.66	4.44	8.05	3.46	4.93		7.08	3.24	12.08	5.11	11.85	10.23	9.74	6.99	2.98	0.53	2.82
L	6.02	1.21	0.15	0.34	15.96	2.01	0.27	24.58	0.59		16.07	1.28	1.59	0.73	0.30	5.10	8.26	11.58	1.23	2.72
M	5.21	0.81	0.08	0.22	9.33	1.67	0.13	16.97	1.34	36.72		0.87	0.82	0.53	0.14	3.68	9.22	8.79	1.90	1.58
N	7.16	0.98	3.33	1.52	3.71	6.52	5.24	3.88	7.10	7.47	5.58		1.64	5.75	0.54	21.27	9.41	2.54	1.12	5.26
P	9.08	1.15	1.71	0.70	10.51	3.32	1.48	8.50	1.80	13.64	5.78	3.15		1.08	0.30	16.15	11.97	4.75	1.15	3.78
Q	2.29	0.71	2.10	7.83	2.62	2.66	6.32	3.30	6.78	16.79	5.17	3.33	2.69		17.22	8.94	4.31	3.00	1.01	2.91
R	3.08	1.00	0.58	1.65	4.62	10.48	11.74	5.14	14.18	5.13	1.88	12.14	0.54	7.43		7.98	3.58	3.76	1.29	3.80
S	22.67	2.73	0.42	0.59	5.63	11.90	0.77	6.21	1.93	9.37	5.26	5.31	1.67	1.97	0.47		15.15	5.08	1.06	1.78
T	14.80	2.56	0.23	0.35	5.15	3.03	0.61	12.57	1.39	17.74	8.67	2.23	1.64	0.47	0.30	16.42		9.67	0.57	1.60
V	6.78	0.76	0.09	0.05	7.13	4.07	0.08	31.23	0.15	20.79	11.00	0.81	1.79	0.14	0.07	4.47	7.86		0.78	1.96
W	3.14	1.25	0.66	1.03	20.65	8.11	0.44	9.88	0.67	16.50	5.00	1.13	0.85	0.60	0.57	6.34	2.87	6.88		13.43
Y	1.93	2.97	1.32	1.76	27.27	0.96	9.42	4.95	2.00	11.23	3.31	7.03	0.59	2.07	0.42	6.93	4.57	3.42	7.85	

c) Dominio matriz

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		0.78	0.19	0.29	2.33	21.26	0.41	7.81	3.49	5.15	3.36	2.37	3.49	2.00	0.39	24.22	14.99	5.33	0.20	1.94
C	2.92		0.08	0.34	7.44	7.15	5.73	1.39	0.55	11.25	1.24	2.52	0.89	0.71	0.68	14.53	15.24	1.18	3.52	22.63
D	0.82	0.06		40.58	0.41	0.57	12.53	2.59	1.77	1.01	0.42	19.29	1.13	2.01	1.69	3.58	9.97	0.94	0.06	0.55
E	2.27	0.39	7.61		2.65	2.29	1.51	3.69	6.70	14.06	2.79	12.80	7.65	8.01	1.73	6.45	6.88	10.28	0.27	1.98
F	4.89	0.60	0.15	0.34		0.99	1.42	12.11	0.30	38.69	6.10	1.52	0.33	0.43	0.11	2.61	3.26	4.19	0.87	21.08
G	10.12	0.93	0.88	1.34	4.84		0.41	2.57	3.42	4.59	4.05	14.74	10.22	0.81	2.65	14.54	10.84	5.19	6.63	1.24
H	2.48	0.32	9.48	1.54	3.85	3.60		2.16	9.23	6.80	1.76	16.30	3.92	12.85	0.90	10.68	6.21	1.01	0.78	6.14
I	6.93	0.92	0.19	0.28	5.57	1.14	0.45		2.41	29.00	7.37	1.97	2.73	1.05	0.45	5.84	12.05	18.44	0.27	2.93
K	2.33	0.40	5.07	8.40	2.12	5.53	2.35	2.76		4.52	2.61	14.95	7.01	10.09	9.14	9.73	8.15	2.06	0.31	2.49
L	3.97	0.70	0.26	0.82	11.80	3.83	1.78	17.30	3.38		12.28	4.64	2.11	2.58	0.98	8.65	9.94	8.93	2.27	3.76
M	4.97	0.40	0.17	0.47	4.88	0.68	0.78	21.82	3.02	24.96		1.67	0.90	1.65	1.11	6.43	11.54	11.31	1.84	1.41
N	5.31	0.51	5.75	3.08	2.81	7.38	10.14	3.13	11.43	4.49	2.15		2.92	3.79	4.21	17.36	8.64	1.91	0.23	4.75
P	9.96	0.46	1.81	3.23	3.41	2.91	1.31	4.91	3.52	8.63	3.30	7.02		4.47	0.60	19.49	12.70	4.19	5.68	2.38
Q	1.95	0.55	1.41	9.78	4.62	3.86	4.93	2.78	7.37	8.86	4.01	7.09	7.68		15.37	8.22	4.80	2.48	1.55	2.70
R	0.96	1.68	1.40	1.83	5.07	10.12	3.94	3.10	10.17	5.82	2.38	20.96	0.67	5.72		7.24	1.69	2.99	1.03	13.22
S	12.93	0.76	2.27	2.56	2.93	10.31	2.11	3.76	4.51	9.28	4.27	11.17	5.23	1.84	0.76		18.93	2.99	0.32	3.05
T	10.69	0.55	1.21	1.07	5.12	3.77	2.18	9.53	2.50	12.48	6.78	6.40	5.20	1.71	1.05	20.78		5.98	0.53	2.46
V	9.15	0.44	0.22	0.37	4.58	2.98	0.25	23.01	4.38	28.42	4.39	1.30	3.31	0.31	0.71	6.53	5.45		0.48	3.72
W	1.42	0.45	0.19	0.29	17.30	1.28	6.05	5.95	2.11	16.02	4.34	2.27	0.94	1.27	14.99	3.53	4.90	3.34		13.35
Y	1.08	0.41	5.34	1.71	35.24	0.64	14.29	4.07	4.21	8.22	3.39	3.72	0.99	2.87	5.56	3.10	2.16	2.61	0.38	

Tabla 12. Frecuencias relativas para cada una de las 19 substituciones posibles de cada aminoácido salvaje. Los valores numéricos son porcentuales. La columna del título representa los aminoácidos salvajes mientras que la fila del título representa a los aminoácidos mutantes a los que puede cambiar.

Previamente, hay que tener en cuenta que las substituciones pueden ser debidas a transiciones o transversiones en el DNA. También podría ocurrir que un cambio de aminoácido se debiera a dos mutaciones a la vez en el mismo triplete de nucleótidos, aunque este suceso es muy improbable (Tabla 13). Así, algunas mutaciones ocurren más frecuentemente que otras y se ha observado que en el mtDNA animal existe un exceso de transiciones sobre transversiones (Keller et al., 2007).

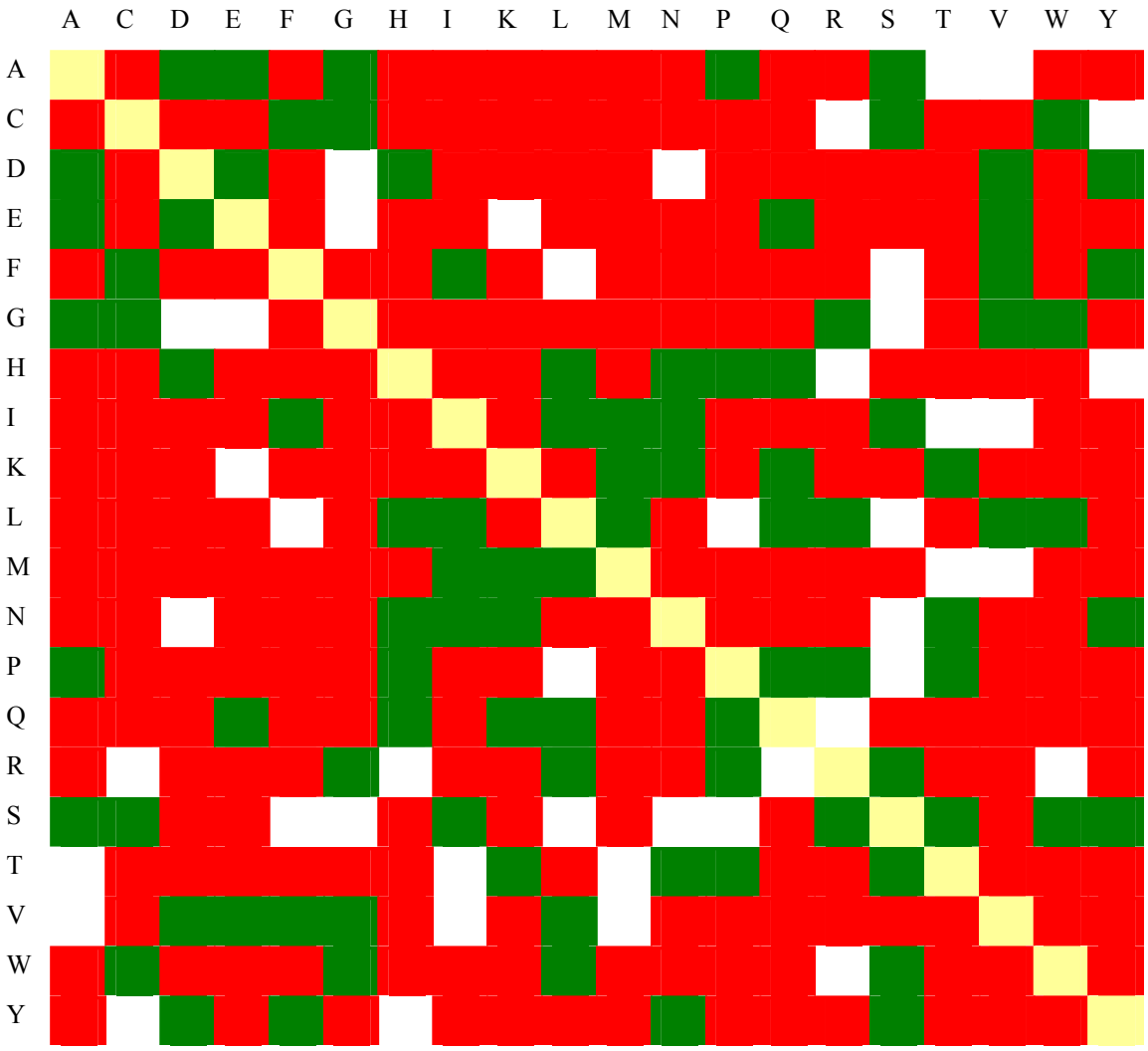

Tabla 13. Matriz de sustituciones con código de colores remarcando si la sustitución se debe a una transición (blanco), una transversión (verde) o una doble sustitución (rojo). Las mutaciones sinónimas (transiciones o transversiones) que no provocan cambio de aminoácido se indican en amarillo. La columna del título representa los aminoácidos salvajes mientras que la fila del título representa a los aminoácidos mutantes a los que puede cambiar.

Sorprendentemente. existen muchos ejemplos de variantes causadas por transversiones que, en contra de lo esperable, son más frecuentes que otras originadas por transiciones. A modo de ejemplo, podemos citar el cambio de glicina (G) a ácido glutámico (E) o ácido aspártico (D) en el dominio transmembrana. Ambos son transiciones y sin embargo, su frecuencia es de 0.3 %, mucho más baja que la presentada por el cambio de glicina (G) a alanina (A), una transversión con 34,5 % de frecuencia de aparición. Esto sugiere que aunque las substituciones de aminoácidos debidas a transiciones debieran ser más frecuentes, el efecto bioquímico de los nuevos aminoácidos tendrá importancia a la hora de determinar si pueden observarse en las

secuencias, dado que la selección negativa eliminará aquellas variantes cuyas propiedades físicoquímicas no sean permisibles.

Por otro lado, podría pensarse que si las transiciones son más frecuentes, una transición que diera lugar a un aminoácido con una frecuencia muy baja a través de la evolución sería susceptible de resultar patológica y su baja aparición podría explicarse por actuación de selección negativa. Revisando las cuatro mutaciones patológicas descritas en el dominio intermembrana, se observa que dos de ellas (p.M1T en MT-ATP6 y p.M1T en MT-CO2) son debidas a la transición con frecuencia más elevada en dicho dominio dentro de todos los cambios posibles del aminoácido metionina (bien es cierto que la posición afectada es la posición inicial del polipéptido y una mutación en ella puede evitar el inicio de la traducción de la molécula). Sin embargo, las otras dos variantes patológicas, p.L135P en MT-CO2 y p.Y278C en MT-CYB, son debidas a transiciones que dan lugar a aminoácidos muy infrecuentes. En ambos casos, los aminoácidos que se derivan de estas transiciones son menos frecuentes que algunos de los derivados de transversiones (Tabla 12).

También hemos analizado la frecuencia de un mismo cambio en diferentes dominios. A priori, si en un determinado dominio una sustitución en particular se encuentra muy poco representada con relación a los otros dos dominios, podría significar que dicho cambio afecta a la estabilidad funcional-estructural (Tabla 14). Así, aparecen cambios muy poco favorecidos como el A-P o T-P en el dominio transmembrana que se sabe que afectan a las alfa hélices predominantes en dicho dominio, el cambio de carga eléctrica K-E o la introducción de carga N-D también en el dominio transmembrana.

CAMBIO	IM	TM	M
A-P	4,4	0,8	1,8
H-D	0,4	0,8	3,7
K-E	3,3	0,7	4,3
N-D	8,7	1,6	3,1
Q-R	1,0	6,0	6,5
T-P	2,4	0,9	3,4
W-R	0,2	0,1	3,2
Y-D	0,1	0,4	2,9

Tabla 14. Cambios con diferencia de frecuencias relativas más relevante (coloreados en rojo) en alguno de los dominios. IM, TM y M son abreviaturas de intermembrana, transmembrana y matriz, respectivamente.

Para concluir el estudio de este discriminador se compararon los valores medios obtenidos para el total de mutaciones de mdmv.1 de cada dominio, tanto para el grupo de substituciones patológicas como neutras. La comparativa (Tabla 15), demuestra que las diferencias en el valor del discriminador son estadísticamente significativas (test Wilcoxon de suma de rangos) con un nivel de significancia de 0,05 para el caso del dominio transmembrana pero no para las otros dos dominios. Esto es así seguramente por el hecho de que el número de mutaciones patológicas de dichos dominios es muy reducido (4 en el intermembrana y 11 en la matriz), no pudiendo por ello, establecerse una comparativa realista.

Discriminador 3	total	intermembrana	transmembrana	matriz
Mutaciones Patológicas	9,4	7,9	9,3	10,07
Mutaciones neutras	12,7	12,6	13,3	11,22
Valor p	0,0003	0,3319	0,0009	0,1904

Tabla 15. Comparativa por dominios del valor medio del discriminador 3 obtenido para el grupo de mutaciones patológicas y mutaciones neutras de la base de datos mdmv.1.

6.5.4. Discriminador descartado: Frecuencia polimórfica del aminoácido mutante en humanos

El análisis de la frecuencia poblacional en humanos de una substitución en una posición determinada es un criterio clásico utilizado ampliamente en el análisis de la patogenicidad de una mutación. Una substitución frecuente no debería ser patológica y una mutación patológica no debería estar presente en la población sana.

Sin embargo hemos desestimado el uso de este discriminador para nuestro clasificador por presentar valores anómalos tanto para el grupo de mutaciones patológicas (algunas posiciones afectadas con frecuencias demasiado elevadas) como para el de mutaciones neutras (substituciones con frecuencias muy bajas). A modo de ejemplo, existen mutaciones patológicas en mdmv.1 con valores altos para este parámetro (4 ‰ en p.R340H de MT-ND4 y 1,6 ‰ en p.M64V de MT-ND6) y por otro lado, un total de 1600 mutaciones neutras presentan un valor inferior a 0.1 ‰. (fichero anexo 5.xls). El enorme número de mutaciones neutras con frecuencias polimórficas excesivamente bajas hace inviable el uso de esta propiedad para la correcta discriminación de mutaciones de clase patológica y neutra.

Las razones de estos sesgos pueden ser varias. Por un lado, el tamaño de las muestras utilizadas como poblaciones control es generalmente muy pequeño y estas poblaciones están además, no muy bien caracterizadas desde un punto de vista clínico, lo que provoca que, algunas veces, incluyan individuos con mutaciones patológicas. Esto es más común cuando las mutaciones presentan penetrancia incompleta y requieren de la contribución de otros factores para manifestar su patogenicidad (López-Gallardo et al., 2014). Según esta hipótesis existirían mutaciones neutras que en realidad podrían ser patológicas. Esto explicaría la presencia de mutaciones neutras con valores tan bajos de frecuencia polimórfica. Por otro lado, en la base de datos GenBank puede encontrarse una sobrerrepresentación de secuencias humanas con presencia de algunas mutaciones patológicas que han sido ampliamente analizadas en los últimos años debido a que están claramente asociadas a mitocondriopatías. Para estas posiciones el valor de este discriminador sería alto sin ser realmente un polimorfismo poblacional.

En nuestro estudio con aproximadamente 30000 secuencias de polipéptidos humanos de cada gen, sólo 9 (15,8 %) de las 57 mutaciones patológicas descritas en mdmv.1 muestran una frecuencia superior a 0,1 ‰. Estas mutaciones concretas incluyen 7 que se han asociado frecuentemente a fenotipos bien conocidos como el síndrome de

Leigh de herencia materna (MILS), la neuropatía sensitiva con ataxia y retinitis pigmentaria (NARP) (Thorburn and Rahman, 1993) o la neuropatía óptica hereditaria de Leber (LHON) (Yu-Wai-Man and Chinnery, 1993). Además, algunas mutaciones asociadas a LHON muestran penetrancia incompleta.

Seguramente en el futuro, cuando la base de datos GenBank incluya un número mucho más alto de secuencias humanas del mtDNA y estas sean realmente representativas de la población mundial, el valor de este parámetro sea por fin interesante y aplicable en entrenamientos de este tipo de predictores.

6.6. Evaluación de Mitoclass.1 y comparación con otros predictores

Hemos comparado nuestro clasificador con algunos de los predictores más ampliamente utilizados: Polyphen-2 (con la opción HumDiv como modelo de clasificación), Provean (con las opciones que aparecen por defecto en el programa) y Mutpred. Para Polyphen-2 tanto las mutaciones catalogadas como "possibly damaging" como las denominadas "probably damaging" se han considerado patológicas en nuestro estudio comparativo. En el caso de Mutpred, hemos utilizado los resultados previamente descritos para mutaciones no sinónimas en el mtDNA (Pereira et al., 2011) con un punto de corte de 0,75 para diferenciación de variantes patológicas.

La razón por la cual hemos considerado la utilización de estos tres predictores es, aparte de su prestigio, el hecho de que permiten el análisis al mismo tiempo de un listado amplio de mutaciones (batch submission). Otros predictores que podrían haberse incorporado en el análisis comparativo sólo permitían la predicción de una variante cada vez, haciendo su uso inviable en el estudio de miles de substituciones como en nuestro caso.

Los resultados comparativos (fichero anexo6.xls) se han obtenido utilizando una base de datos de validación para evitar el sesgo debido a la inclusión de mutaciones utilizadas en el entrenamiento del predictor. Esta base de datos de validación se ha creado a partir de la base de datos completa mdmv.1 excluyendo el 60 % de mutaciones seleccionadas para el entrenamiento (training dataset) (fichero anexo7.xls). Los parámetros evaluados han sido elegidos por ser los más ampliamente utilizados en la comparativa entre este tipo de predictores.

6.6.1. Resultados de Polyphen-2, Provean y Mutpred sobre la base de datos mdmv.1 y la base de datos de validación

Antes de efectuar la comparativa de Mitoclass.1 frente a los otros predictores utilizando exclusivamente la base de datos de validación hemos comprobado los resultados de los tres predictores para la base de datos completa mdmv.1 (fichero anexo1.xls). El análisis completo (Tabla 16) muestra que Polyphen-2 es el predictor con mejor sensibilidad de los tres (94,7 %) con sólo tres falsos negativos. Por el contrario, Mutpred ha mostrado una sensibilidad muy baja (57,9 %) con 24 falsos negativos de las 57 mutaciones clasificadas como patológicas. Provean ha obtenido resultados intermedios con una sensibilidad del 87,7 % y 7 falsos negativos. Respecto a la especificidad, se observan diferencias entre Polyphen-2 y Provean, con resultados más favorables para Provean (59,2 % frente a 46,9 %).

El análisis de la base de datos de validación muestra por su parte idénticos resultados de sensibilidad para Provean y Polyphen-2 (91,3 %) con una pequeña ventaja en especificidad para Provean (60.3 % frente a 47,7 % de Polyphen-2) que ya había sido observada al analizar la base de datos completa mdmv.1 (Tabla 17). Respecto a Mutpred, podemos concluir que un punto de corte de 0,75 tal y como recomiendan los autores ((Li et al., 2009), no ha resultado interesante para el cribado de variantes no sinónimas del mtDNA.

Estos resultados muestran que aunque frente a la base de datos de validación Provean obtuvo una ligera ventaja sobre Polyphen-2, al evaluar la base de datos completa de 2835 variantes de mdmv.1 (Tabla 16) se aprecia una mejor sensibilidad para Polyphen-2. Considerando que la finalidad del proyecto es conseguir un test de screening mejorado sobre otros clasificadores disponibles, podríamos concluir que Polyphen-2 es el predictor con mejores prestaciones. Esto siempre considerando clasificar las predicciones "probably damaging" y "possibly damaging" como patológicas y las predicciones "benign" como neutras y teniendo en cuenta además su especificidad mejorable y un pequeño número de variantes para las cuales dicho predictor no logra generar resultados ("unknown").

	POLYPHEN-2*	PROVEAN	MUTPRED
Sensibilidad	94,7	87,7	57,9
Especificidad	46,9	59,2	87,3
TP	54	50	33
FN	3	7	24
TN	1303	1646	2426
FP	1475	1132	352

Tabla 16. Comparativa entre predictores con la base de datos completa mdmv.1 de 2835 mutaciones (57 patológicas y 2778 neutras). TP, TN, FP, FN son las abreviaturas de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

* Polyphen-2 no ha permitido la predicción de 25 variantes debido, según autores, a que una parte de la secuencia inicial y final de p.MT-ND5 no está bien alineada por regiones con repeticiones o mucha variación en la composición de aminoácidos de dichas regiones. Para poder incluir estas variantes en la comparación las hemos considerado como predicciones neutras.

6.6.2. Resultados obtenidos por Mitoclass.1 en la validación frente al resto de predictores

Tras evaluar dichos predictores podemos establecer los objetivos de un clasificador que mejore las características de Polyphen-2 y los resultados que debería lograr en la validación:

a) Una sensibilidad ≥ 91,3 %.

b) Una especificidad ≥47,7 % para intentar reducir el número de falsos positivos.

c) Una predicción para el 100 % de variantes consultadas sin resultados "unknown".

d) Una clasificación dicotómica de las predicciones (patológicas/neutras) para evitar tomas de decisión subjetivas como las empleadas con Polyphen-2 debido al hecho de que clasifica los resultados en tres grupos (probably damaging, possibly damaging y benign).

Nuestro clasificador Mitoclass.1 cumple estas premisas, con una sensibilidad de 95,6 % sobre la base de datos de validación (comparada con un 91,3 % de Provean y Polyphen-2) y una especificidad de 51,5 %, ligeramente por encima de la de Polyphen-2 (47,7 %). Además, Mitoclass.1 (como Provean), logra predecir el 100 % de las variantes analizadas a diferencia de Polyphen-2 que clasifica 10 mutaciones de p.MT-ND5 como "unknown". Este efecto se repite fuera de la base de datos de validación con un total de

25 substituciones para las cuales Polyphen-2 no genera resultados como hemos comentado previamente.

Finalmente, hemos evaluado el área bajo la curva (AUC) de las curvas ROC para los cuatro predictores sobre la base de datos de validación (Figura 16). Los resultados son muy similares para todos ellos al igual que los obtenidos para el MCC (Mathews correlation coefficient) (Tabla 17).

	MITOCLASS.1	**POLYPHEN-2***	**PROVEAN**	**MUTPRED**
Sensibilidad	95,65	91,30	91,30	60,87
Especificidad	51,53	47,73	60,35	85,61
AUC	82,47 %	81,00 %	79,96 %	80,38 %
TP	22	21	21	14
FN	1	2	2	9
TN	555	514	650	922
FP	522	563	427	155
MCC	0,14	0,11	0,15	0,18

Tabla 17. Comparativa entre predictores con la base de datos de validación de 1100 mutaciones (23 patológicas y 1077 neutras). AUC se refiere a área bajo la curva mientras que MCC se refiere a Mathews correlation coefficient. TP, TN, FP, FN son las abreviaturas de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

* Polyphen-2 no ha permitido la predicción de 10 variantes debido, según autores, a que una parte de la secuencia inicial y final de p.MT-ND5 no está bien alineada por regiones con repeticiones o mucha variación en la composición de aminoácidos de dichas regiones. Para poder incluir estas variantes en la comparación las hemos considerado como predicciones neutras.

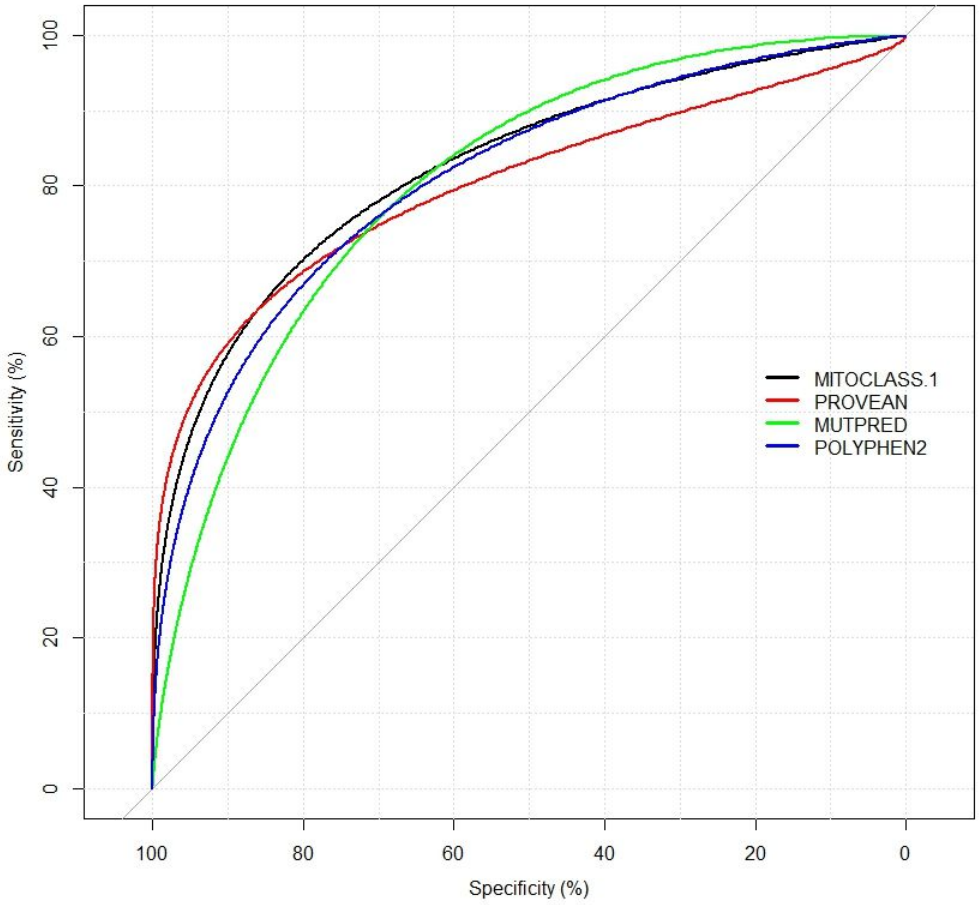

Figura 16. Curva ROC para los cuatro predictores generada sobre los resultados predictivos de la base de datos de validación.

6.6.3. Análisis de los falsos negativos obtenidos por los predictores evaluados en la etapa de validación

Analizando los falsos negativos obtenidos en el grupo de validación se puede observar que tanto Provean como Polyphen-2 clasifican como neutra una mutación correspondiente a la transición m.10158T>C (p.S34P en p.MT-ND3). Esta mutación muestra una relación inversa en cíbridos de osteosarcoma 143B entre carga mutacional y actividad del complejo 1 (McFarland et al., 2004a). Además, la mutación se ha descrito en varios pedrigís (Bugiani et al., 2004; Crimi et al., 2004; Lebon et al., 2003; McFarland et al., 2004a) y su patogenicidad está bien documentada (Mitchell et al., 2006). Por otro lado, la transición m.3700G>A (p.A132T en p.MT-ND1) ha sido clasificada como neutra únicamente por Polyphen-2. Esta mutación está descrita como

una mutación primaria rara de neuropatía óptica hereditaria de Leber(Achilli et al., 2012).

Además de estas dos variantes descritas, Provean no clasifica correctamente como patológica la transversión m.4171C>A (p.L289M en p.MT-ND1), una mutación primaria de neuropatía óptica hereditaria de Leber (Kim et al., 2002).

Estas tres mutaciones afectan a posiciones con bajas conservaciones en eucariotas (fichero anexo6.xls). Sin embargo, Mitoclass.1 logra una correcta predicción para todas ellas, debido a que el predictor no utiliza la conservación de una posición concreta como discriminador. A través del discriminador 1, que es una suma tanto de la conservación como de la coevolución de la posición, logramos que mutaciones en posiciones poco conservadas pero con signos interesantes de coevolución puedan ser predichas como patológicas. La única mutación patológica presente en el grupo de validación que Mitoclass.1 no consigue clasificar correctamente es p.V65A en p.MT-ND4L. La razón es que se trata de un polimorfismo muy común en eucariotas con un valor para el discriminador 2 de 35 % (Tabla 18).

Variante	Predictor	CI	D1	D2	D3
p.S34P p.MT-ND3	Polyphen-2 y Provean	10,15	60,53	0,34	5,23
p.V65A p.MT-ND4L	Mitoclass.1	24,72	103	35,06	6,78
p.L289M p.MT-ND1	Provean	29,72	78,64	1,99	16,07
p.A132T p.MT-ND1	Polyphen-2	72,82	123,98	0,30	12,59

Tabla 18. Análisis de los atributos discriminadores para las predicciones de falsos negativos en Provean, Polyphen-2 y Mitoclass.1 en el grupo de mutaciones del validation dataset. CI se refiere al índice de conservación, mientras que D1, D2 y D3 son las abreviaturas de los tres atributos discriminadores utilizados por Mitoclass.1.

6.6.4. Análisis de los falsos positivos obtenidos por los predictores evaluados en la etapa de validación

Finalmente, también hemos analizado los resultados falsos positivos, observando que de las 1077 mutaciones neutras incluidas en el validation dataset, 328 (30,4 % del total) son clasificadas correctamente como neutras por los cuatro métodos evaluados. Sin embargo, 122 mutaciones neutras son clasificadas como patológicas por todos ellos (11,3 % del total). Sólo 84 mutaciones (7 % del total) son identificadas como patológicas únicamente por Mitoclass.1 Esto parece indicar que la elevada sensibilidad de Mitoclass.1 no resulta en un incremento excesivo de falsos positivos exclusivos de nuestro test. Analizando ese 7 % de variantes, se comprueba que en su mayoría, son mutaciones poco conservadas en eucariotas. Muchas de ellas son clasificadas como patológicas por sus valores numéricos para el discriminador 2 (aminoácidos mutantes muy infrecuentes) y/o altos valores del discriminador 1 (signos de coevolución).

6.7. Análisis de mutaciones con evidencias dudosas sobre su verdadera patogenicidad

Existe un pequeño número de mutaciones neutras (37) del validation dataset que fueron clasificadas por Mitomap como mutaciones patológicas en el grupo "mtDNA Mutations with Reports of Disease-Associations". Sin embargo, nosotros no hemos encontrado evidencias robustas suficientes que permitan su clasificación como patológicas (siguiendo nuestro criterio de patogenicidad descrito en material y métodos). Por ello, las hemos incluido en el grupo de mutaciones neutras.

Para este conjunto de variantes, hemos analizado los resultados predictivos de los cuatro métodos comparados, considerando que sería importante que un predictor no descartara esas substituciones y poder ser así posteriormente confirmadas en el laboratorio. La finalidad de estos predictores es servir de cribado inicial para filtrar aquellas mutaciones que no muestren indicios de patogenicidad por lo que es interesante que dispongan de buena sensibilidad aun a pesar de presentar falsos positivos y poder, más tarde, confirmarlos por métodos experimentales como el uso de cíbridos. Por ello, hemos analizado la especificidad de los cuatro predictores para este grupo de variantes, observando que el número de falsos positivos es siempre superior al índice de falsos positivos general del grupo de validación. Esto indica que algunas de estas

substituciones podrían ser realmente patológicas y que la realización de pruebas confirmatorias complementarias sería adecuado para terminar de definir su patogenicidad. La especificidad de Mitoclass.1, Polyphen-2, Provean y Mutpred para este grupo es de 35,1 %, 43 %, 45,9 % y 70 % respectivamente, comparada con la del grupo de validación de 51,5 %, 47 %, 60,3 % y 85,6 % respectivamente. La mayor diferencia en especificidad aparece para Mitoclass.1 resultando en un mayor número de falsos positivos que el resto para este grupo concreto. Esto podría ser indicativo de una mejor discriminación de potenciales mutaciones patológicas en favor de Mitoclass.1. Ocho de dichas mutaciones (21,6 %) fueron clasificadas como neutras por todos los predictores mientras que 10 variantes (27,0 %) resultaron patológicas por todos los métodos evaluados. Los resultados demuestran que Mitoclass.1 clasificó como patológicas 24 de las 37, tres más que Polyphen-2 y cuatro más que Provean (fichero anexo8.xls).

Como hemos comentado, diez mutaciones fueron clasificadas como patológicas por los cuatro clasificadores (Tabla 19B). Para ellas, hemos comparado los valores medios de los tres atributos discriminadores respecto al conjunto de mutaciones neutras incluidas en el grupo de validación (1077 variantes). La comparación muestra diferencias (discriminador 1=124,1 % vs. 80,0 %, discriminador 2=0,6 % vs. 7,9 % y discriminador 3=11,0 % vs. 12,7 %). Por lo tanto, parece que la confirmación por el resto de predictores analizados de un resultado patológico ofrecido por Mitoclass.1 refuerza la posibilidad de que se trate de una verdadera mutación patológica.

Por otro lado, cuatro mutaciones fueron consideradas patológicas únicamente por Mitoclass.1 (Tabla 19A). Para ellas, hemos observado que los valores numéricos de los atributos discriminadores están más próximos a los valores medios de las mutaciones neutras del validation dataset. Por ello, dichas predicciones no muestran la misma seguridad que las anteriores, aunque sí es cierto que los valores reflejan cierta tendencia hacia la patogenicidad, haciendo que Mitoclass.1 las considere como patológicas (discriminador 1=92,4 % vs. 80,0 %, discriminador 2=4,1 % vs. 7,9 % y discriminador 3=17,1 % vs. 12,7 %).

Mutación rCRS	AA subs/PP/Dom	D1	D2	D1
A	Valores medios del grupo A	92,48	4,16	17,10
m.9738G>T	p.A178S/p.MT-CO3/TM	97,94	0,41	21,06
m.9972A>C	p.I256L/p.MT-CO3/IM	112,25	0,97	25,85
m.3475G>A	pA147T/p.MT-ND1/TM	84,35	2,32	12,59
m.8557G>A	pA11V/p.MT-ATP6/TM	75,40	12,97	8,93
B	Valores medios del grupo B	124,15	0,67	11,06
m.3407G>A	p.A55G/p.MT-ND2/TM	127,50	2,32	12,66
m.4160T>C	p.F60S/p.MT-ND2/TM	113,68	0,01	4,22
m.5244G>A	p.G259S/p.MT-ND2/TM	138,87	1,00	23,65
m.10543A>G	p.H25R/p.MT-ND4L/TM	150,47	2,56	0,66
m.13051G>A	p.G239S/p.MT-ND5/TM	163,88	0,01	23,65
m.13511A>T	p.K392M/p.MT-ND5/TM	98,85	0,01	3,24
m.15243G>A	p.G166E/p.MT-CYB/IM	115,60	0	11,92
m.8668T>C	p.W48R/p.MT-ATP6/M	131,03	0,16	14,99
m.8795A>G	p.H90R/p.MT-ATP6/TM	73,77	0	0,66
m.8528T>C	p.W55R/p.MT-ATP8/M	102,63	0,04	14,99
	Valores medios variantes neutras del validation dataset	80,05	7,90	12,70

Tabla 19. Valores de los discriminadores 1-3 para las substituciones con evidencias dudosas sobre su patogenicidad.

Grupo A) Mutaciones únicamente consideradas como patológicas por Mitoclass.1. Grupo B) Mutaciones consideradas como patológicas por los cuatro predictores.

Mutación rCRS, AA subs, PP, Dom, D1, D2 y D3 son las abreviaturas de mutación acorde a la secuencia de referencia "revised Cambridge Reference Sequence", substitución de aminoácidos, polipéptido, dominio y valores numéricos para los discriminadores 1-3, respectivamente.

6.8. Predicción de todas las posibles variantes no sinónimas para los trece polipéptidos codificados por el DNA mitocondrial humano

En nuestro trabajo también presentamos los resultados para las 24201 variantes posibles en los trece polipéptidos codificados por el mtDNA humano (sin incluir las mutaciones nonsense). En genética, una mutación sin sentido (nonsense) es un tipo de mutación puntual en una secuencia de DNA que provoca la aparición de un codón de terminación prematuro en el RNAm transcrito, lo cual conduce a su vez a la producción de un producto proteico truncado, incompleto y por lo general no funcional.

La secuencia de referencia (rCRS, NC_012920.1) ha sido utilizada para definir los aminoácidos de referencia de cada gen. Los resultados muestran que 16656 substituciones (69,8 %) podrían ser patológicas (fichero anexo9.xls). Cuando comparamos con los resultados predictivos de Polyphen-2 (fichero anexo9.xls), observamos que presenta 149 resultados no concluyentes (clasificados como "unknown") y 18887 mutaciones patológicas (78,0 %). Esta tendencia de Polyphen-2 en cuanto a la predicción de un gran número de mutaciones como patológicas (número alto de falsos positivos) ya lo habíamos detectado previamente durante la evaluación de los tests con el grupo de validación y fue comentado como uno de los puntos débiles del método, al igual que la aparición de predicciones "unknown" no concluyentes. 14792 variantes (61,1 %) han sido clasificadas como patológicas tanto por Mitoclass.1 como por Polyphen-2.

Las predicciones patológicas de Mitoclass.1 no se acumulan en genes concretos. A pesar de que el 77,2 % de las mutaciones confirmadas como patológicas y presentes en mdmv.1 afectan a cuatro genes (*MT-ND1*, *ND5*, *ND6*, *ATP6*), únicamente un 31,6 % de las mutaciones predichas como patológicas afecta a dichos genes (porcentaje muy similar al que se esperaría considerando su número de aminoácidos: 34,9 %) (Tabla 20).

Este sesgo en la composición de la base de datos mdmv.1 podría indicar que fenotipos asociados a mutaciones en algunos genes del mtDNA (*MT-CO1, MT-CO2, MT-CO3, MT-ATP8*,...) no son fácilmente reconocibles como mitocondriopatías y por tanto, se subestima la presencia de mutaciones patológicas en el mtDNA para ellos.
Cuando analizamos la predicción de mutaciones patológicas por dominio con Mitoclass.1, se observa que no tienden a acumularse en ninguno de los tres dominios.

Esto contradice lo que aparece en la base de datos mdmv.1 en la que las mutaciones confirmadas como patológicas se encuentran sobrerrepresentadas en el dominio transmembrana (73,7 % de las variantes totales frente al 61,1 % que resultaría teniendo en cuenta el número de aminoácidos del dominio). Este porcentaje de 61,1 % es similar al obtenido por las predicciones de Mitoclass.1 para dicho dominio. Por otro lado, el número de mutaciones patológicas predichas para el dominio matriz y el intermembrana es de nuevo similar al esperado (Tabla 21).

Complejo	Polipéptido	AA	%	MUT Confirmadas en mdmv.1	%	MUT Predicción Mitoclass.1	%
CI		2214	55,8	36	63,2	8348	50,12
	p.MT-ND1	318	8,4	15	26,3	1441	8,65
	p.MT-ND2	347	9,2	1	1,8	1231	7,39
	p.MT-ND3	115	3,0	2	3,5	489	2,94
	p.MT-ND4	459	12,1	3	5,3	1956	11,74
	p.MT-ND4L	98	2,6	1	1,8	406	2,44
	p.MT-ND5	603	15,9	7	12,3	2358	14,16
	p.MT-ND6	174	4,6	7	12,3	467	2,80
CIII		380	10,0	2	3,5	1823	10,95
	p.MT-CYB	380	10,0	2	3,5	1823	10,95
CIV		1001	26,4	4	7,0	5323	31,96
	p.MT-CO1	513	13,5	1	1,8	2875	17,26
	p.MT-CO2	227	6,0	2	3,5	1081	6,49
	p.MT-CO3	261	6,9	1	1,8	1367	8,21
CV		294	7,8	15	26,3	1162	6,98
	p.MT-ATP6	226	6,0	15	26,3	1000	6,00
	p.MT-ATP8	68	1,8	0	0	162	0,97

Tabla 20. Porcentaje de mutaciones patológicas confirmadas en mdmv,1 y predichas por Mitoclass.1 por polipéptido/complejo. AA, % MUT, %, confirmadas en mdmv,1, predicción Mitoclass.1, se refieren respectivamente a: número de aminoácidos y su porcentaje en un polipéptido particular, número de mutaciones patológicas y su porcentaje en un polipéptido particular, mutaciones patológicas confirmadas en mdmv.1, mutaciones patológicas predichas por Mitoclass,1.

Dominio	AA	%	MUT Confirmadas en mdmv.1	%	MUT Predicción Mitoclass.1	%
	3889	100	57	100	16656	100
IM	747	19,2	4	7,0	3234	19,4
TM	2376	61,1	42	73,7	10179	61,1
M	766	19,7	11	19,3	3243	19,5

Tabla 21. Porcentaje de mutaciones patológicas confirmadas en mdmv,1 y predichas por Mitoclass.1 por dominio. AA, % MUT, %, confirmadas en mdmv.1, predicción Mitoclass.1, IM, TM y M se refieren respectivamente a: número de aminoácidos y su porcentaje en un dominio particular, número de mutaciones patológicas y su porcentaje en un dominio particular, mutaciones patológicas confirmadas en mdmv.1, mutaciones patológicas predichas por Mitoclass.1 y dominios intermembrana, transmembrana y matriz.

6.9. Análisis del grado de coevolución entre residuos de los polipéptidos

Aprovechando el uso de la coevolución en uno de los discriminadores del clasificador Mitoclass.1 decidimos profundizar en la predicción de interacciones entre parejas o grupos de aminoácidos de los polipéptidos en base a la coevolución. El estudio se realizó utilizando algunos programas bioinformáticos de uso frecuente por la comunidad científica en este campo.

6.9.1. Control de calidad de resultados predictivos de coevolución entre aminoácidos

La caracterización de dominios realizada en nuestro proyecto es interesante para la verificación de parejas o grupos de aminoácidos que puedan presentar coevolución. En nuestro trabajo utilizamos programas informáticos como PSICOV (Jones et al., 2012), H2r (Merkl and Zwick, 2008) o MISTIC (Simonetti et al., 2013), para la identificación de estos residuos. Una posible coevolución entre un residuo del dominio matriz y un residuo del dominio intermembrana sería difícil de explicar al encontrarse separados por todo el espacio transmembrana. Sin embargo, una potencial coevolución entre dos aminoácidos que se encuentran en el mismo dominio sería un resultado interesante a tener en cuenta. Por ello, es importante conocer las posiciones que delimitan unos dominios de otros.

A modo de ejemplo, utilizando el programa MISTIC (con los parámetros fijados por defecto), hemos analizado en 5164 polipéptidos ortólogos de p.MT-ND1 las 15 parejas de aminoácidos que presentan mayor probabilidad de coevolución. En cuatro parejas, una de las posiciones se ubica en el dominio intermembrana y la otra en el dominio matriz, haciendo difícil explicar su coevolución. Sin embargo, en las 11 parejas restantes, los aminoácidos para los que el programa sugiere coevolución pertenecen al mismo dominio o a dominios contiguos (Tabla 22).

Posición 1	Aminoácido 1	Dominio 1	Posición 2	Aminoácido 2	Dominio 2
200	L	IM	284	Q	TM
249	A	M	251	S	TM
39	V	M	223	F	TM
118	W	TM	217	A	IM
171	H	IM	255	Y	M
1	M	TM	2	P	TM
148	I	TM	165	L	IM
118	W	TM	255	Y	M
6	L	TM	156	M	TM
255	Y	M	264	L	TM
42	P	M	88	P	IM
88	P	IM	235	N	TM
148	I	TM	301	L	TM
165	L	IM	255	Y	M
39	V	M	200	L	IM

Tabla 22. Las 15 parejas de aminoácidos del polipéptido p.MT-ND1 con mayor grado de coevolución, de acuerdo al programa MISTIC. Cuatro de ellas (marcadas en rojo) muestran posiciones separadas por la membrana interna mitocondrial. IM, M y TM son las abreviaturas de dominio intermembrana, matriz y transmembrana respectivamente.

6.9.2. Análisis de pares de residuos con interacción espacial dentro del mismo POLIPEPTIDO a través del estudio de la coevolución con PSICOV

Antes de seleccionar el indicador cMI proporcionado por el programa MISTIC como parte del discriminador 1 de Mitoclass.1, realizamos diversos estudios previos para corroborar la idoneidad de la coevolución como posible marcador de patogenicidad. Uno de ellos fue la utilización del programa PSICOV (Jones et al,, 2012) para predecir pares de residuos que coevolucionan y que se encuentran interaccionando de forma directa en los polipéptidos codificados por el mtDNA humano. Obtener información sobre si dos residuos poco conservados del mismo polipéptido están en

contacto físico en la estructura tridimensional y muestran coevolución puede resultar relevante para el estudio de la patogenicidad de una mutación.

Para ello, partimos de los alineamientos múltiples generados sobre una base de datos de más de 4000 secuencias ortólogas de especies diferentes para los trece polipéptidos como ficheros de entrada de PSICOV. De todas las predicciones generadas para cada polipéptido se consideraron las L/5 primeras como las de más fiabilidad según indicación de los autores. A modo de ejemplo, se muestran las predicciones para p.MT-CYB (Tabla 23). Las predicciones para los trece polipéptidos se encuentran disponibles en la carpeta suplementaria "psicov_completos".

Las predicciones obtenidas fueron verificadas analizando la distancia interatómica entre los carbonos beta de los residuos del par (o carbonos alfa en el caso de glicina) utilizando el modelo cristalino (PDB id = 1QCR, chain C) perteneciente a *Bos taurus* y considerando una distancia de 8 Å como indicativa de interacción espacial directa, tal y como se recomienda en la literatura (Manavalan and Ponnuswamy, 1978).

Elegimos revisar exclusivamente las predicciones pertenecientes al espacio transmembrana por ser el dominio con estructura secundaria mejor definida, al tratarse de alfa hélices. Hay que considerar que las distancias entre los residuos no han sido medidas directamente en la especie humana por lo que cabría esperar ligeras diferencias. Puede observarse (Tabla 24) que la conservación media de la posición 1 y la posición 2 de todos los pares predichos es muy similar (alrededor del 50 %), lo cual también es un indicador de que la coevolución puede ser real. Por otro lado, la distancia interatómica media ha resultado de 8,1 Å, reforzando la hipótesis de existencia de interacción espacial. De hecho, de las 28 predicciones, 15 muestran distancias inferiores o iguales a 8 Å. Además, el valor predictivo positivo (PPV; positive predictive value) determinado por PSICOV para dichas predicciones es en todos los casos superior a 0,73. Estos resultados parecen indicar que el uso de PSICOV para predecir pares de residuos con contacto espacial a través del análisis de la coevolución es un método que, aun lejos de ser perfecto, permite disponer de información interesante para el análisis de la patogenicidad de una mutación en aquellas posiciones para las que no se observa una alta conservación evolutiva.

POSICIÓN 1	POSICIÓN 2
5	20
8	111
14	19
15	20
17	197
17	23
25	216
27	224
29	230
38	140
39	232
43	82
45	190
47	82
49	184
54	68
56	176
57	176
61	172
68	73
81	248
81	243
84	251
89	235
94	123
98	120
101	106
105	313
107	308
108	308
109	203
109	313
112	200
114	302
117	302

POSICIÓN 1	POSICIÓN 2
118	302
121	295
121	298
121	299
122	192
123	189
124	277
125	150
125	185
125	295
126	185
129	181
135	163
150	160
150	164
158	238
177	264
190	323
226	370
233	324
241	246
248	254
266	345
267	342
269	340
276	294
276	297
276	336
277	294
277	336
318	373
318	374
323	333
327	357
331	354
331	357
332	362
334	350
334	354
338	350
338	351

Tabla 23. Predicción de pares de aminoácidos con posibilidad de coevolución (L/5 predicciones) utilizando el alineamiento múltiple del polipéptido p.MT-CYB. El código de colores representa blanco para el dominio transmembrana, verde para el matriz y amarillo para el intermembrana.

Pos 1	CI Pos 1	Pos 2	CI Pos 2	Distancia (Å)
122	7	192	41	9,7
81	27	243	12,6	5,6
118	47	302	10,4	9,7
43	14,5	82	45,7	8,8
117	15,4	302	10,4	7
123	20,4	189	69,3	7,3
332	24,1	362	17,9	11,4
89	25,3	235	9,2	7,8
129	26,7	181	65,9	7,4
39	3,1	232	10,9	7,4
94	33,4	123	20,4	8,7
334	46,6	354	65,9	7
334	46,6	350	70,5	8,8
331	47,4	357	52,5	6,9
331	47,4	354	65,9	5,7
125	53,2	185	71,1	8,9
125	53,2	295	45,2	12,6
121	58,8	295	45,2	10,3
121	58,8	299	74,9	8,7
114	66,8	302	10,4	6,4
98	71,3	120	78,2	9
49	77,1	184	25,4	10,6
112	79,9	200	97,8	5,2
338	81,6	351	92,5	6,17
338	81,6	350	70,5	8
45	89,6	190	53,8	9,7
126	92,5	185	71,1	6,3
47	93,4	82	45,7	7,5
MEDIA	**49,6**		**48,2**	**8,1**

Tabla 24. Distancia interatómica entre los aminoácidos de cada par predicho por PSICOV para el dominio transmembrana del polipéptido p.MT-CYB dentro de las L/5 primeras predicciones. El alineamiento múltiple de proteínas ortólogas se ejecutó sobre 4134 secuencias. Pos 1 y Pos 2 simbolizan la Posición 1 y la Posición 2 del par de aminoácidos considerados, CI simboliza el índice de conservación de cada posición.

6.9.3. Análisis de pares de residuos con interacción espacial dentro del mismo DOMINIO de cada polipéptido con PSICOV

Por otro lado, revisando las predicciones generadas por PSICOV para p.MT-CYB (Tabla 23) puede observarse que existe un 34,2 % de parejas de aminoácidos (26 de las 76) en las que las posiciones no pertenecen al mismo dominio. Aunque es cierto que estas interacciones pueden ser reales por el propio dinamismo de las proteínas o por encontrarse los residuos en una situación periférica entre dominios, es igualmente válido pensar que puedan tratarse de predicciones erróneas por encontrarse en posiciones situadas en dominios distintos. Predicciones entre posiciones del dominio intermembrana y del dominio matriz, que sí serían francamente extrañas y difíciles de explicar, no han aparecido. De todos modos, con la intención de eliminar este conjunto de predicciones posiblemente anómalas entre posiciones de dominios distintos se consideró fusionar todos los segmentos de cada polipéptido pertenecientes al mismo dominio a partir de los alineamientos múltiples. De esa manera, ejecutamos PSICOV sobre las tres secuencias generadas artificialmente, una para cada dominio de los trece polipéptidos.

Previamente se evaluó si el orden en que los segmentos de cada dominio eran fusionados para crear esta secuencia artificial influía en las predicciones, Para ello se analizó el resultado de PSICOV sobre la secuencia creada con los fragmentos transmembrana de p.MT-CYB diseñándola de tres maneras distintas, uniendo las 8 hélices del dominio (denominadas 1,2,3,4,5,6,7,8) en diferente orden: 12345678, 87654321 y 18273645. Los resultados obtenidos demuestran que la forma en que se fusionan los fragmentos apenas influye en las predicciones, De las 39 predicciones L/5, los resultados son idénticos para 12345678 y 18273645 y solo hay dos resultados diferentes para 87654321 (fichero anexo10.doc).

Para valorar la ventaja de utilizar estas secuencias artificiales en lugar del polipéptido completo, hemos comparado los resultados de PSICOV para el polipéptido p.MT-CYB utilizando el polipéptido completo y considerando los dominios por separado. De las 75 predicciones obtenidas al usar las secuencias artificiales, observamos que 37 (49,3 %) no han sido incluidas entre las L/5 predicciones obtenidas al usar PSICOV con el alineamiento del polipéptido completo. La mayor cantidad de predicciones comunes aparece para el espacio transmembrana por ser el más representado. Sin embargo, el hecho de separar los dominios intermembrana y matriz ha

permitido localizar muchas parejas con posible coevolución en el interior de dichos dominios que al analizar la proteína completa no eran detectadas o no estaban incluidas dentro de las L/5 primeras predicciones de mayor fiabilidad por su reducido PPV (positive predictive value) (fichero anexo11.doc).

Estos resultados demuestran que tal vez sea interesante considerar ambas tácticas: las predicciones tanto del polipéptido completo como de las secuencias fusión de dominio. De esta manera podríamos aumentar el número de predicciones de los dominios con menor número de aminoácidos (intermembrana y matriz) y utilizar el hecho de que la predicción sea común al analizarla con ambas metodologías como un criterio para asignarle mayor fiabilidad.

Sin embargo, aunque a priori parezca interesante, al comprobar la calidad de las predicciones del dominio transmembrana de p.MT-CYB que no han sido detectadas utilizando el alineamiento completo pero si el alineamiento de las secuencias artificiales de dominio (columna PREDICCION COMUN = "N" en fichero anexo11.doc) hemos observado que la distancia interatómica media es de 16 Å (Tabla 25) con varios pares de aminoácidos teóricamente bastante alejados en el espacio (comparado con 8,1 Å de distancia interatómica media de las predicciones con el polipéptido completo para ese dominio). Esto parece indicar que la calidad de las predicciones para los dominios por separado es inferior que la del polipéptido completo. De hecho, si revisamos los valores de PPV para las L/5 predicciones de los trece polipéptidos separados por dominios se observan valores en su mayoría bajos (PPV < 0,5) para un número muy importante de predicciones. Los resultados se muestran en el fichero anexo12.xls.

Pos 1	CI Pos 1	Pos 2	CI Pos 2	Distancia (Å)
38	85	197	83,5	10,3
238	2,8	327	20,4	26,2
225	36,6	326	93,4	30,1
365	50,7	370	21	8,6
121	58,8	298	82,7	8,2
102	69,5	304	15,3	8,5
308	71,1	355	73,2	21,2
49	77,1	183	90,4	7,8
37	94,2	93	85	6
100	95,9	177	97,7	35,4
Total	**64,17**		**66,26**	**16,23**

Tabla 25. Distancia interatómica entre los aminoácidos de cada par predicho por PSICOV para el dominio transmembrana del polipéptido p.MT-CYB utilizando la secuencia fusión artificial de dominio para las L/5 primeras predicciones. La tabla muestra exclusivamente las predicciones no detectadas al utilizar el alineamiento del polipéptido completo. El alineamiento múltiple de proteínas ortólogas se ejecutó sobre 4134 secuencias, Pos 1 y Pos 2 simbolizan la Posición 1 y la Posición 2 del par de aminoácidos considerados. CI simboliza el índice de conservación de cada posición.

6.9.4. Análisis de las interacciones determinadas por PSICOV para las mutaciones patológicas de la base de datos mdmv.1 con baja conservación interespecífica

Hemos analizado los resultados al aplicar PSICOV tanto con alineamientos de proteínas completas como con alineamientos específicos de dominio (ya detallado en el apartado previo) para aquellas posiciones en las que existen mutaciones patológicas incorporadas en la base de datos mdmv.1 y que además presentan conservación media-baja. La posibilidad de que dicha posición coevolucione con otra podría postularse como una de las causas de su patogenicidad. Una mutación en una de las dos posiciones podría romper la interacción entre ambas y provocar problemas en la funcionalidad o estructura de la proteína. Para corroborar este planteamiento revisamos las interacciones espaciales predichas por PSICOV en el conjunto de mutaciones patológicas con CI menor de 80 % presentes en mdmv.1 (19 mutaciones). De ellas, hemos encontrado indicios de coevolución con PSICOV (Tabla 26) para 9 de ellas (47,3 %).

Gen	Posición del aa	CI (%)	posición covariante		
MT-ATP6	155	67,29	103 (23) DOM	95 (87) DOM	202 (89) COM
MT-ATP6	222	71,37	112 (60) COM		
MT-ND1	289	29,72	308 (60) DOM	115 (62) DOM	280 (75) COM
MT-ND1	52	36,61	83 (72) COM	57 (25) COM	
MT-ND3	45	14,31	51 (95) DOM	22 (66) COM	
MT-ND5	312	79,42	254 (64) COM		
MT-ND6	36	15,94	29 (15) DOM		
MT-ND6	63	75,87	169 (60) COM	164 (11) COM	
MT-ND6	60	77,84	18 (22) COM		

Tabla 26. Parejas de aminoácidos con posible coevolución según PSICOV para posiciones donde se han descrito mutaciones no sinónimas con conservación media-baja (CI menor de 80 %) en los genes codificantes del mtDNA incluidos en mdmv.1. CI=conservation index. En la columna "Posición covariante" se muestra la posición del aminoácido que coevoluciona, entre paréntesis el CI de dicha posición y en mayúsculas si la predicción de PSICOV se ha obtenido utilizando el alineamiento de la proteína completa (COM) o por dominios (DOM). Las predicciones sombreadas en verde son las que presentan valores de CI más próximos al CI de la posición con la que coevolucionan y por ello, a priori, más probables.

En nuestro análisis, hemos decidido dar mayor validez a aquellas parejas en las que las dos posiciones muestran conservación parecida, descartando de este modo predicciones como las de las posiciones 52-83 de p.MT-ND1 al presentar valores de CI de 36 % y 72 % respectivamente (Tabla 26). De entre las predicciones restantes eliminamos la pareja 222-112 de p.MT-ATP6 al tratarse de residuos de la matriz y espacio intermembrana, siendo por ello, improbable la interacción física entre ambos.

De este modo, habría cuatro mutaciones patológicas que podrían explicarse por rotura de interacciones directas y que vamos a verificar estudiando la proximidad espacial de los residuos.

a) La interacción 312-254 de p.MT-ND5 fue corroborada analizando la estructura cristalina de la proteína homóloga nqo12 en *Thermus thermophilus* (código PDB: 4HEA chain L). El resultado demuestra que las posiciones pertenecen a hélices alfa distintas del espacio transmembrana. Dichas hélices están enfrentadas y la interacción es factible. La distancia entre los residuos medida en la proteína homóloga ha resultado de 11,9 Å (Figura 17).

Figura 17. Distancia entre los residuos 305-247 de la proteína nqo12 de *Thermus thermophilus* (posiciones homólogas a 312-254 de p.MT-ND5 en Homo sapiens).

b) La interacción 36-29 en p.MT-ND6 se produce entre dos posiciones localizadas en la misma hélice del espacio transmembrana y a una distancia de 10,4 Å medida en la proteína homóloga nqo10 de *Thermus thermophilus* (código PDB: 4HEA chain J). Sin embargo, ambas posiciones se encuentran separadas por un giro de hélice y existen residuos entre ellas dificultando por ello la interacción física directa (Figura 18).

Figura 18. Distancia entre los residuos 36-29 de la proteína nqo10 de *Thermus thermophilus* (posiciones homólogas a 36-29 de p.MT-ND6 en Homo sapiens).

c) La interacción 63-169 en p.MT-ND6 parece más extraña ya que analizando de nuevo la estructura de la proteína homóloga en *Thermus thermophilus* (código PDB: 4HEA chain J), uno de los residuos se encuentra en mitad de una hélice del espacio transmembrana mientras que el otro está localizado en la matriz.

d) La interacción 52-57 en p.MT-ND1 también podría ser correcta, Ambos residuos se encuentran en la matriz y analizando la estructura cristalina de nqo8 en *Thermus thermophilus* (código PDB: 4HEA chain H), la distancia entre residuos es de 9,2 Å (Figura 19). Esta mutación en la posición 52 es una transición (m.3460G>A) que convierte alanina en treonina en p.MT-ND1 (Howell et al., 1991; Huoponen et al., 1991). En nuestro panel de organismos de 5165 especies aparece conservada tan sólo en un 36,5 % de ellas. Una posible coevolución con treonina en la posición 57 explicaría la importancia funcional de una posición tan poco conservada. Además, para confirmar la coevolución de estos residuos se estudió la presencia del par A52-T57 en el árbol filogenético creado con las 5165 especies apareciendo en dos ramas independientes: Protostomos y Deuterostomos (Mammalia) (Figura 20). Los protóstomos son una agrupación de filos del Reino Animal que junto con los deuteróstomos, forman los dos grandes linajes en que se dividen los bilaterales (Bilateria).

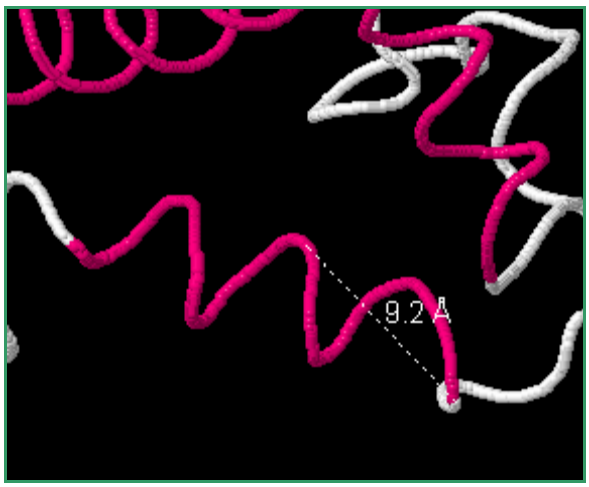

Figura 19. Distancia entre los residuos 63-68 de la proteína nqo8 de *Thermus thermophilus* (posiciones homólogas a 52-57 de p.MT-ND1 en Homo sapiens).

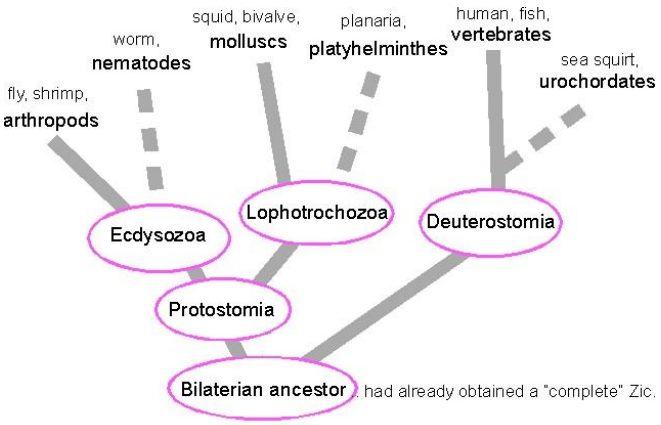

Figura 20. Árbol filogenético utilizado para verificar que la interacción p.A52-T57 en p.MT-ND1 aparece en dos ramas diferentes (Protostomia y Deuterostomia).

Extraida de : http://www,med,nagasaki-u,ac,jp/phrmch1/lcn/tool_kit_evolution,htm.

De esta manera, podríamos concluir que, en al menos dos de las mutaciones patológicas de posiciones poco conservadas, es probable que exista una interacción física definida por coevolución y postular esto como causa de su patogenicidad (52-57 de p.MT-ND1 y 312-254 de p.MT-ND5).

6.9.5. Identificación de redes de interacción de residuos coevolutivos del mismo polipéptido con el programa H2r

El programa H2r analiza, a diferencia de PSICOV, residuos con signos de coevolución sin necesidad de que exista una interacción espacial directa entre ellos. Los resultados aportados por este programa son por tanto, complementarios a los ofrecidos por PSICOV ya que residuos alejados pueden estar conectados de forma coevolutiva y resultar importantes para la función o estructura de la proteínas. Los resultados completos para los trece genes se encuentran disponibles en el fichero suplementario anexo13.xls.

A modo de ejemplo, las dos mutaciones patológicas de p.MT-ATP6 con conservación media baja (en las posiciones 155 y 222) ya habían sido determinadas por PSICOV como posiciones con signos de coevolución aun a pesar de que al analizar sus parejas coevolutivas se comprobó que no existía interacción física evidente entre ellas. Sin embargo, estas dos posiciones vuelven a aparecen como importantes con el programa H2r (Tabla 27). Esto podría indicar que dichas posiciones forman parte de

redes funcionales-estructurales importantes junto con otros aminoácidos del polipéptido que no tienen que estar necesariamente cerca en el espacio.

Residuo	Posición	Residuo	Posición
S	8	I	138
A	11	L	141
P	12	L	149
Q	56	A	155
W	68	I	164
Q	97	T	165
T	112	L	196
V	113	L	222
L	129	H	223

Tabla 27. Residuos con signos de coevolución según el programa H2r para p.MT-ATP6. Sombreadas en verde se muestran las dos posiciones patológicas de la base de datos mdmv.1 con conservación media-baja (CI menor de 80 %).

Para las posiciones patológicas con conservación media-baja (CI menor de 80 %) del resto de polipéptidos no hemos encontrado ningún resultado positivo con H2r a excepción de la posición 36 de p.MT-ND6 (Tabla 28). Curiosamente, dicha posición también había sido descrita como interesante por PSICOV y la habíamos descartado porque la pareja de interacción se encontraba separada por un giro de hélice alfa. Así pues, las tres posiciones identificadas por H2r también lo habían sido por PSICOV aunque sus parejas coevolutivas resultaron encontrarse a más de 8 Å. Esto justificaría el uso de H2r para descubrir interacciones sin contacto físico. De este modo, 5 de las 19 mutaciones patológicas con conservación media-baja podrían teóricamente afectar a posiciones con rasgos interesantes de coevolución que explicaran su carácter deletéreo.

6.9.6. Identificación de signos de coevolución entre polipéptidos de un mismo complejo respiratorio con el programa H2r

Los trece polipéptidos codificados por el mtDNA forman parte del sistema de fosforilación oxidativa de la mitocondria y se integran dentro de los complejos OXPHOS junto con otros polipéptidos de origen nuclear. Por ello, sería interesante que el análisis de la coevolución entre residuos no solamente abarcara a residuos

pertenecientes al mismo polipéptido. Lo ideal sería analizar coevolución entre residuos de los polipéptidos pertenecientes al mismo complejo. Sin embargo, evaluar la presencia de coevolución utilizando alineamientos múltiples de las proteínas de origen nuclear no es todavía realista debido al reducido número de homólogos presentes actualmente en las bases de datos. Por ello, el estudio debe centrarse en el análisis de los polipéptidos de origen mitocondrial incluidos en un mismo complejo. Así pues, analizamos por un lado las interacciones coevolutivas del complejo I (polipéptidos p.MT-ND1, ND2, ND3, ND4, ND4L, ND5, ND6), complejo IV (p.MT-CO1, CO2, CO3) y complejo V (p.MT-ATP6, ATP8).

La presencia de coevolución fue evaluada con el programa H2r para favorecer la búsqueda de interacciones alejadas en el espacio. Para ello se generaron las proteínas "fusión" a través de los alineamientos múltiples uniendo todos los ficheros de alineamientos de polipéptidos de un mismo complejo en uno único siguiendo el orden comentado en el párrafo anterior.

Los resultados para las mutaciones confirmadas como patológicas presentes en posiciones con conservación media-baja de la base de datos mdmv.1 muestran coevolución entre polipéptidos diferentes para 6 de las 19 casos. De ellas, dos interacciones no habían aparecido previamente ni al utilizar PSICOV ni H2r sobre los polipéptidos por separado y podrían dar explicación a la patogenicidad de dichas posiciones. Se trata de la posición 289 de p.MT-ND1 y de la posición 72 de p.MT-ND6 (fichero anexo13.xls).

En el primer caso aparece una interacción de la posición 289 de p.MT-ND1 con la posición 18 de p.MT-ND6. Ambas posiciones se encuentran en el espacio transmembrana y se sabe que ambos polipéptidos están muy próximos. En el segundo, la posición 72 de p.MT-ND6 muestra coevolución con las posiciones 247 y 307 de p.MT-ND5, todas ellas pertenecientes también al espacio transmembrana. En este caso, la interacción es más difícilmente explicable, puesto que p.MT-ND1 y p.MT-ND5 están bastante alejados en la estructura cristalina del complejo I. De todos modos, es posible que los complejos sean dinámicos y que posiciones alejadas puedan acercarse durante el proceso de la fosforilación oxidativa (Figura 21).

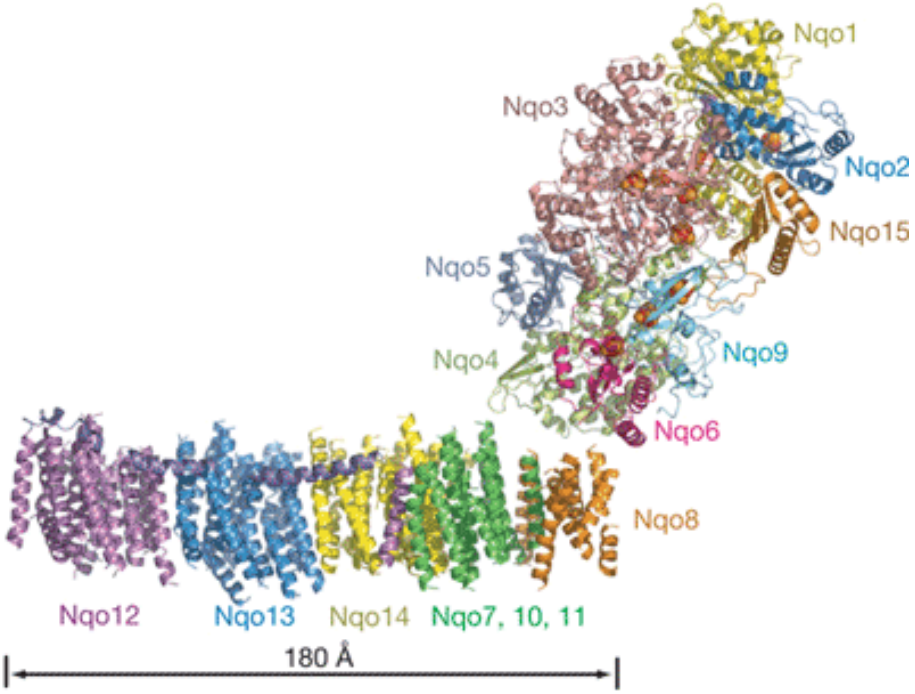

Figura 21. Estructura cristalina del complejo I determinada en Thermus thermophilus. Nqo8 corresponde a p.MT-ND1, Nqo10 a p.MT-ND6 y Nqo12 a p.MT-ND5 en humanos (Efremov et al., 2010).

6.9.7. Conclusiones sobre el estudio de coevolución

De esta manera, al concluir nuestro estudio sobre interacciones dependientes de coevolución (PSICOV para interacciones directas en polipéptidos independientes, PSICOV sobre posiciones del mismo dominio y polipéptido, H2r sobre polipéptidos independientes y H2r sobre complejos respiratorios), hemos detectado presencia de la misma en 10 de las 19 mutaciones patológicas confirmadas presentes en la base de datos mdmv.1 pertenecientes a posiciones con conservación media-baja. Toda esta información, a pesar de no formar parte de los discriminadores usados en Mitoclass.1, debido a que necesitábamos un parámetro cuantificable numéricamente (obtenido con el score cMI del programa MISTIC), pensamos que era importante incluirla en la tesis como parte final de los resultados.

El estudio de coevolución con PSICOV y H2r es un buen complemento para conocer más en detalle la naturaleza de las interacciones y los residuos afectados. Finalmente, hay que tener presente que los complejos se asocian en supercomplejos, por lo que podrían darse más ejemplos de coevolución a un nivel todavía superior y esto podría estudiarse en trabajos futuros. Además, el análisis de la coevolución con

programas bioinformáticos todavía está en pleno proceso de desarrollo y tienen cabida avances importantes en el futuro (de Juan et al,, 2013).

GEN	Posición	Predicción Coevolución
MT-ATP6	155	H2r - PSICOV
MT-ATP6	222	H2r - PSICOV
MT-ND1	2	--
MT-ND1	289	H2R complejos - PSICOV
MT-ND1	132	--
MT-ND1	128	--
MT-ND1	52	PSICOV
MT-ND2	71	--
MT-ND3	34	--
MT-ND3	45	PSICOV
MT-ND4L	65	--
MT-ND5	312	PSICOV
MT-ND6	64 *	--
MT-ND6	36	H2r - PSICOV
MT-ND6	72	H2r complejos
MT-ND6	25	--
MT-ND6	63	PSICOV
MT-ND6	60	PSICOV

Tabla 28. 19 mutaciones patológicas de la base de datos mdmv,1 con conservación media-baja (CI menor de 80 %) y sus resultados para PSICOV o H2r (para polipéptidos independientes o para los polipéptidos del mismo complejo unidos: "H2r complejos"). Las mutaciones para las que no ha aparecido predicción muestran "--" en la columna "Predicción Coevolución".

* La posición 64 de p.MT-ND6 alberga dos substituciones patológicas distintas.

7. Conclusiones

El objetivo principal del proyecto ha sido conseguido y fruto de este trabajo surge Mitoclass.1, un predictor bioinformático de patogenicidad de mutaciones no sinónimas en el mtDNA. Mitoclass.1 presenta unos niveles de sensibilidad ligeramente superiores a otros programas ampliamente utilizados como Polyphen-2, Provean o Mutpred así como una especificidad mejorada respecto a Polyphen-2, programa considerado como referencia mundial en la actualidad.

Para la consecución del objetivo principal fue necesario desarrollar previamente una serie de objetivos secundarios que hicieron posible diseñar el clasificador.

a) Nuestros resultados confirman que la curación de las bases de datos existentes es un requisito imprescindible para poder entrenar correctamente al predictor y poder evaluar después prestaciones como sensibilidad o especificidad del método. Un análisis de dichas bases de datos nos permitió descubrir mutaciones clasificadas como patológicas que en realidad no lo son.

Ante la importancia de este hecho, hemos elaborado la primera base de datos revisada de mutaciones no sinónimas en el mtDNA humano a la que hemos denominado mdmv.1. Las mutaciones contenidas en dicha base de datos han sido clasificadas como neutras o patológicas de acuerdo a una serie de criterios de patogenicidad que en nuestro laboratorio hemos considerado como requisitos para definir una mutación como patológica.

b) Algunos atributos muy razonables y ampliamente utilizados en el estudio de la patogenicidad de este tipo de mutaciones no han resultado convenientes para su aplicación en el entrenamiento del predictor. En casos como el de la frecuencia poblacional en humanos de una determinada mutación apareció un número muy elevado de mutaciones supuestamente neutras con valores muy bajos para este indicador, cuando lo esperable sería lo contrario.

Por otra parte, tras desestimar varios atributos, descubrimos que tan solo la combinación de tres discriminadores sencillos permitían obtener el clasificador que nos habíamos propuesto en nuestro proyecto. Estos atributos se basaban en propiedades como la

conservación de la posición en organismos eucariotas y el grado de coevolución que presentaban con otras posiciones del polipéptido.
Además, dado que las proteínas estudiadas son en todos los casos proteínas integrales de membrana y expuestas a tres condiciones ambientales diferentes, se incorporó como discriminador la frecuencia de aparición evolutiva de los tipos de aminoácidos mutantes en cada uno de los tres dominios presentes en los polipéptidos. Incluir este atributo permitió mejorar la predicción. Antes de eso fue necesaria una caracterización por dominios de cada uno de los trece polipéptidos, tampoco realizada previamente, utilizando estructuras cristalinas de especies homólogas.

c) En nuestro estudio también comprobamos la importancia de elegir correctamente el algoritmo de clasificación. Utilizando la aplicación Weka, que permite valorar las diferencias entre distintos métodos de clasificación, verificamos que el algoritmo bayesiano ingenuo ofrecía los mejores resultados.
Además, tuvimos que salvar el inconveniente de contar con una base de datos desbalanceada. El clasificador se entrenó utilizando la base de datos mdmv.1 diseñada por nuestro grupo. Sin embargo, debido al escaso número de mutaciones patológicas publicadas y la incidencia que la sobrerrepresentación de mutaciones neutras podía tener sobre el entrenamiento del predictor y la sensibilidad del método, se ejecutó el algoritmo SMOTE para generar mutaciones patológicas sintéticas e igualar el número de mutaciones patológicas y neutras.

8. Consideraciones futuras

El hecho de ejecutar el entrenamiento del predictor exclusivamente con variantes curadas procedentes del mtDNA, así como una cuidadosa selección de los discriminadores más adecuados ha permitido que Mitoclass.1 alcance unos buenos niveles de sensibilidad y especificidad. A pesar de ello, pensamos que en el futuro el predictor podrá optimizarse y lograr prestaciones aún mejores utilizando para ello la nueva información que rápidamente va incorporándose en las bases de datos como GenBank o Mitomap.

Por un lado, el número de secuencias de referencia publicadas de proteínas homólogas de especies distintas a la humana aumenta rápidamente. Esto posibilitará que los cálculos de conservación y coevolución en los que están basados los discriminadores del clasificador sean en el futuro mucho más representativos de la situación real.

Por otro lado, el número de mutaciones humanas candidatas a resultar patológicas identificadas por secuenciación crece continuamente. De esta manera, la base de datos mdmv.1 curada por nuestro grupo en este trabajo también podría ser actualizada periódicamente permitiendo reentrenar al predictor con la nueva información.

Todo esto es factible debido a la facilidad con la que el predictor puede actualizarse al usar herramientas bioinformáticas fácilmente automatizables.

9. Listado de archivos suplementarios

Los archivos suplementarios se encuentran disponibles en un fichero comprimido en la siguiente dirección http://biblos.unizar.es/zaguan/tesis/TUZ_950_anexos.zip

- anexo1.xls

  Base de datos mdmv.1 con la predicción (patológica/neutra) de cada una de las 2835 mutaciones clasificadas con los tres programas evaluados (Polyphen-2, Provean y Mutpred). La tabla incluye las puntuaciones (scores) obtenidas por cada predictor para cada mutación, así como su clasificación dicotómica (patológica/neutra) o tricotómica para el caso de Polyphen-2.

- anexo2.xls

  índice de conservación (CI) de todas las posiciones de los trece polipéptidos codificados por el mtDNA humano medido en secuencias ortólogas eucariotas. La tabla incluye el número de secuencias ortólogas de cada gen utilizadas para el cálculo.

- anexo3.xls

  índice de conservación (CI) de todas las posiciones de los trece polipéptidos codificados por el mtDNA humano medido exclusivamente en secuencias ortólogas bacterianas. La columna "homólogos bacterianos" incluye en el título el número de secuencias ortólogas utilizadas en el cálculo. La columna "recuento de aminoácidos en dicha posición" presenta el número de aminoácidos de cada tipo presentes en el alineamiento múltiple en la misma columna que la posición del polipéptido humano considerada.

- anexo4.xls

  Valor numérico del discriminador 2 utilizado por el predictor Mitoclass.1 para cada una de las mutaciones de la base de datos mdmv.1. Los dominios 1,2 y 3 se refieren respectivamente a intermembrana, transmembrana y matriz.

- anexo5.xls

  Frecuencia polimórfica humana (medida en tanto por mil) para cada una de las mutaciones presentes en la base de datos mdmv.1. La tabla incluye una columna con el número de secuencias humanas utilizadas para el cálculo.

- anexo6.xls

  Resultados comparativos de Mitoclass.1, Polyphen-2, Provean y Mutpred para la base de datos de validación. La tabla incluye una columna con el valor del CI para cada una de las mutaciones, así como el valor numérico de los tres discriminadores usados en Mitoclass.1.

- anexo7.xls

  Mutaciones seleccionadas para el grupo de entrenamiento y de validación, procedentes de la base de datos mdmv.1.

- anexo8.xls

  Resultados de los predictores evaluados para las mutaciones presentes en el grupo de validación clasificadas como neutras en la base de datos mdmv.1 pero con ciertos signos de posible patogenicidad según la web Mitomap. La tabla incluye una columna con el recuento de predicciones patológicas para cada mutación considerando los cuatro predictores (Mitoclass.1, Polyphen-2, Provean y Mutpred).

- anexo9.xls

  Resultados predictivos de Mitoclass.1 y Polyphen-2 para el total de posibles mutaciones no sinónimas en los trece polipéptidos codificados por el mtDNA humano. Los dominios 1,2 y 3 se refieren respectivamente a intermembrana, transmembrana y matriz.

- anexo10.doc

  Predicción de interacciones directas con el programa PSICOV para la secuencia "dominio" transmembrana de p.MT-CYB. La secuencia dominio fue diseñada ordenando las 8 hélices alfa de tres formas diferentes como muestran la columna 1, 2 y 3. En cada fila aparecen dos valores, indicativos del número de residuo del polipéptido para cada par. Se muestran en amarillo las dos predicciones distintas para la secuencia 87654321. Una de ellas no aparece en las otras dos variantes mientras que la otra sí aparece en ellas pero no en la 87654321.

- anexo11.doc

  Comparativa entre los pares de residuos predichos por PSICOV (primeras L/5 predicciones) utilizando las secuencias "dominio" o el polipéptido completo. Los resultados muestran el análisis de p.MT-CYB. La columna "predicción común" indica una "N" de "no" para los pares no detectados por PSICOV al utilizar el polipéptido completo. El código de colores simboliza: verde para el dominio matriz, blanco para el transmembrana y amarillo para el intermembrana.

- anexo12.xls

  Predicciones de interacciones directas con el programa PSICOV por dominios (transmembrana, matriz e intermembrana) para cada uno de los trece polipéptidos codificados por el mtDNA humano. Una de las columnas de cada predicción es el valor PPV ofrecido por el programa (Positive predictive value). Pos 1 y Pos 2 corresponden con la posición 1 y la posición 2 del polipéptido que formarían la interacción.

- anexo13.xls

  Predicciones con el programa H2R para interacciones transitivas de residuos con coevolución. En la tabla figura una columna con las predicciones teniendo en cuenta únicamente el polipéptido y otra columna teniendo en cuenta todos los polipéptidos del mismo complejo respiratorio. En ambas columnas se representa entre paréntesis el número de secuencias ortólogas utilizadas en el alineamiento múltiple utilizado por el programa.

- carpeta suplementaria "psicov_completos"

  La carpeta contiene el fichero de salida del programa PSICOV para cada uno de los trece genes codificantes del mtDNA humano. Se trata de ficheros de texto con varias columnas. La primera y segunda columnas son las posiciones del polipéptido que coevolucionan. La tercera columna siempre es 0 y la cuarta siempre es 8. Estas dos columnas no son interesantes. La última columna indica el valor predictivo positivo (PPV).

10. Referencias

Acharya, V., and Nagarajaram, H.A. (2012). Hansa: an automated method for discriminating disease and neutral human nsSNPs. Hum. Mutat. *33*, 332–337.

Achilli, A., Iommarini, L., Olivieri, A., Pala, M., Hooshiar Kashani, B., Reynier, P., La Morgia, C., Valentino, M.L., Liguori, R., Pizza, F., et al. (2012). Rare Primary Mitochondrial DNA Mutations and Probable Synergistic Variants in Leber's Hereditary Optic Neuropathy. PLoS ONE *7*, e42242.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457–465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. *23*, 147–147.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Bandelt, H.-J., Achilli, A., Kong, Q.-P., Salas, A., Lutz-Bonengel, S., Sun, C., Zhang, Y.-P., Torroni, A., and Yao, Y.-G. (2005). Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem. Biophys. Res. Commun. *333*, 122–130.

Bandelt, H.-J., Yao, Y.-G., Salas, A., Kivisild, T., and Bravi, C.M. (2007). High penetrance of sequencing errors and interpretative shortcomings in mtDNA sequence analysis of LHON patients. Biochem. Biophys. Res. Commun. *352*, 283–291.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2015). GenBank. Nucleic Acids Res. *43*, D30–D35.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Birrell, J.A., and Hirst, J. (2010). Truncation of subunit ND2 disrupts the threefold symmetry of the antiporter-like subunits in complex I from higher metazoans. FEBS Lett. *584*, 4247–4252.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Bromberg, Y., Yachdav, G., and Rost, B. (2008). SNAP predicts effect of mutations on protein function. Bioinformatics *24*, 2397–2398.

Brown, M.D., Voljavec, A.S., Lott, M.T., Torroni, A., Yang, C.C., and Wallace, D.C. (1992). Mitochondrial DNA complex I and III mutations associated with Leber's hereditary optic neuropathy. Genetics *130*, 163–173.

Brunham, L.R., Singaraja, R.R., Pape, T.D., Kejariwal, A., Thomas, P.D., and Hayden, M.R. (2005). Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLoS Genet. *1*, e83.

Bugiani, M., Invernizzi, F., Alberio, S., Briem, E., Lamantea, E., Carrara, F., Moroni, I., Farina, L., Spada, M., Donati, M.A., et al. (2004). Clinical and molecular findings in children with complex I deficiency. Biochim. Biophys. Acta *1659*, 136–147.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. Publ. Protein Soc. *13*, 190–202.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. *30*, 1237–1244.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinforma. Oxf. Engl. *23*, 1875–1882.

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinforma. Oxf. Engl. *22*, 2729–2734.

Castellana, S., and Mazza, T. (2013). Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. Brief. Bioinform. bbt013.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. ArXiv11061813 Cs.

Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins *62*, 1125–1132.

Chinnery, P.F. (1993). Mitochondrial Disorders Overview. In GeneReviews(®), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J. Bean, T.D. Bird, C.-T. Fong, H.C. Mefford, R.J. Smith, et al., eds. (Seattle (WA): University of Washington, Seattle),.

Chinnery, P.F., Brown, D.T., Andrews, R.M., Singh-Kler, R., Riordan-Eva, P., Lindley, J., Applegarth, D.A., Turnbull, D.M., and Howell, N. (2001). The mitochondrial ND6 gene is a hot spot for mutations that cause Leber's hereditary optic neuropathy. Brain J. Neurol. *124*, 209–218.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS One *7*, e46688.

Clark, K.M., Taylor, R.W., Johnson, M.A., Chinnery, P.F., Chrzanowska-Lightowlers, Z.M., Andrews, R.M., Nelson, I.P., Wood, N.W., Lamont, P.J., Hanna, M.G., et al. (1999). An mtDNA mutation in the initiation codon of the cytochrome C oxidase subunit II gene results in lower levels of the protein and a mitochondrial encephalomyopathy. Am. J. Hum. Genet. *64*, 1330–1339.

Crimi, M., Papadimitriou, A., Galbiati, S., Palamidou, P., Fortunato, F., Bordoni, A., Papandreou, U., Papadimitriou, D., Hadjigeorgiou, G.M., Drogari, E., et al. (2004). A new mitochondrial DNA mutation in ND3 gene causing severe Leigh syndrome with early lethality. Pediatr. Res. *55*, 842–846.

De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. *40*, D935–D939.

Dehouck, Y., Kwasigroch, J.M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics *12*, 151.

Dietrich, S., Borst, N., Schlee, S., Schneider, D., Janda, J.-O., Sterner, R., and Merkl, R. (2012). Experimental assessment of the importance of amino acid positions identified by an entropy-based correlation analysis of multiple-sequence alignments. Biochemistry (Mosc.) *51*, 5633–5641.

DiMauro, S., and Schon, E.A. (2001). Mitochondrial DNA mutations in human disease. Am. J. Med. Genet. *106*, 18–26.

DiMauro, S., and Schon, E.A. (2003). Mitochondrial respiratory-chain diseases. N. Engl. J. Med. *348*, 2656–2668.

Druzhyna, N.M., Wilson, G.L., and LeDoux, S.P. (2008). Mitochondrial DNA repair in aging and disease. Mech. Ageing Dev. *129*, 383–390.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids.

Efremov, R.G., and Sazanov, L.A. (2012). The coupling mechanism of respiratory complex I - a structural and evolutionary perspective. Biochim. Biophys. Acta *1817*, 1785–1795.

Efremov, R.G., Baradaran, R., and Sazanov, L.A. (2010). The architecture of respiratory complex I. Nature *465*, 441–445.

Elson, J.L., Sweeney, M.G., Procaccio, V., Yarham, J.W., Salas, A., Kong, Q.-P., van der Westhuizen, F.H., Pitceathly, R.D.S., Thorburn, D.R., Lott, M.T., et al. (2012). Toward a mtDNA locus-specific mutation database using the LOVD platform. Hum. Mutat. *33*, 1352–1358.

Fan, W., Waymire, K.G., Narula, N., Li, P., Rocher, C., Coskun, P.E., Vannan, M.A., Narula, J., Macgregor, G.R., and Wallace, D.C. (2008). A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. Science *319*, 958–962.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. Proteins *57*, 811–819.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Frank S Cordes, J.N.B. (2002). Proline-Induced Distortions of Transmembrane Helices. J. Mol. Biol. *323*, 951–960.

Friedrich, T., and Böttcher, B. (2004). The gross structure of the respiratory complex I: a Lego System. Biochim. Biophys. Acta *1608*, 1–9.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics *14*, 1–13.

Goldstein, A.C., Bhatia, P., and Vento, J.M. (2013). Mitochondrial disease in childhood: nuclear encoded. Neurother. J. Am. Soc. Exp. Neurother. *10*, 212–226.

Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A., and Tishkoff, S.A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. *24*, 757–768.

Guo, X., Yin, Y., Dong, C., Yang, G., and Zhou, G. (2008). On the Class Imbalance Problem. In Proceedings of the 2008 Fourth International Conference on Natural Computation - Volume 04, (Washington, DC, USA: IEEE Computer Society), pp. 192–201.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl *11*, 10–18.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. *225*, 487–494.

Herráez, A. (2006). Biomolecules in the computer: Jmol to the rescue. Biochem. Mol. Biol. Educ. Bimon. Publ. Int. Union Biochem. Mol. Biol. *34*, 255–261.

Hicks, S., Wheeler, D.A., Plon, S.E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum. Mutat. *32*, 661–668.

Hong, S., and Pedersen, P.L. (2004). Mitochondrial ATP synthase: a bioinformatic approach reveals new insights about the roles of supernumerary subunits g and A6L. J. Bioenerg. Biomembr. *36*, 515–523.

Howell, N., Bindoff, L.A., McCullough, D.A., Kubacka, I., Poulton, J., Mackey, D., Taylor, L., and Turnbull, D.M. (1991). Leber hereditary optic neuropathy: identification of the same mitochondrial ND1 mutation in six pedigrees. Am. J. Hum. Genet. *49*, 939–950.

Huoponen, K., Vilkki, J., Aula, P., Nikoskelainen, E.K., and Savontaus, M.L. (1991). A new mtDNA mutation associated with Leber hereditary optic neuroretinopathy. Am. J. Hum. Genet. *48*, 1147–1153.

Johns, D.R., Neufeld, M.J., and Park, R.D. (1992). An ND-6 mitochondrial DNA mutation associated with Leber hereditary optic neuropathy. Biochem. Biophys. Res. Commun. *187*, 1551–1557.

Johnson, M.A., Bindoff, L.A., and Turnbull, D.M. (1993). Cytochrome c oxidase activity in single muscle fibers: assay techniques and diagnostic applications. Ann. Neurol. *33*, 28–35.

Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinforma. Oxf. Engl. *28*, 184–190.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. *36*, D202–D205.

Keller, I., Bensasson, D., and Nichols, R.A. (2007). Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. PLoS Genet *3*, e22.

Kim, H.J., Khalimonchuk, O., Smith, P.M., and Winge, D.R. (2012). Structure, function, and assembly of heme centers in mitochondrial respiratory complexes. Biochim. Biophys. Acta *1823*, 1604–1616.

Kim, J.Y., Hwang, J.-M., and Park, S.S. (2002). Mitochondrial DNA C4171A/ND1 is a novel primary causative mutation of Leber's hereditary optic neuropathy with a good prognosis. Ann. Neurol. *51*, 630–634.

Kirby, D.M., Salemi, R., Sugiana, C., Ohtake, A., Parry, L., Bell, K.M., Kirk, E.P., Boneh, A., Taylor, R.W., Dahl, H.-H.M., et al. (2004). NDUFS6 mutations are a novel cause of lethal neonatal mitochondrial complex I deficiency. J. Clin. Invest. *114*, 837–845.

Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Kumar, S., Bellis, C., Zlojutro, M., Melton, P.E., Blangero, J., and Curran, J.E. (2011). Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. BMC Evol. Biol. *11*, 293.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A., et al. (2006). Machine learning in bioinformatics. Brief. Bioinform. *7*, 86–112.

Lebon, S., Chol, M., Benit, P., Mugnier, C., Chretien, D., Giurgea, I., Kern, I., Girardin, E., Hertz-Pannier, L., Lonlay, P. de, et al. (2003). Recurrent de novo mitochondrial DNA mutations in respiratory chain deficiency. J. Med. Genet. *40*, 896–899.

Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinforma. Oxf. Engl. *25*, 2744–2750.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinforma. Oxf. Engl. *17*, 282–283.

López-Gallardo, E., Emperador, S., Solano, A., Llobet, L., Martín-Navarro, A., López-Pérez, M.J., Briones, P., Pineda, M., Artuch, R., Barraquer, E., et al. (2014). Expanding the clinical phenotypes of MT-ATP6 mutations. Hum. Mol. Genet. *23*, 6191–6200.

Mackey, D., and Howell, N. (1992). A variant of Leber hereditary optic neuropathy characterized by recovery of vision and by an unusual mitochondrial genetic etiology. Am. J. Hum. Genet. *51*, 1218–1228.

Maechler, P., and de Andrade, P.B.M. (2006). Mitochondrial damages and the regulation of insulin secretion. Biochem. Soc. Trans. *34*, 824–827.

Malfatti, E., Bugiani, M., Invernizzi, F., de Souza, C.F.-M., Farina, L., Carrara, F., Lamantea, E., Antozzi, C., Confalonieri, P., Sanseverino, M.T., et al. (2007). Novel mutations of ND genes in complex I deficiency associated with mitochondrial encephalopathy. Brain J. Neurol. *130*, 1894–1904.

Manavalan, P., and Ponnuswamy, P.K. (1978). Hydrophobic character of amino acid residues in globular proteins. Nature *275*, 673–674.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE *6*, e28766.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. *30*, 1072–1080.

Martelotto, L.G., Ng, C.K., De Filippo, M.R., Zhang, Y., Piscuoglio, S., Lim, R.S., Shen, R., Norton, L., Reis-Filho, J.S., and Weigelt, B. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. *15*, 484.

Martin, L.C., Gloor, G.B., Dunn, S.D., and Wahl, L.M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*, 4116–4124.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA - Protein Struct. *405*, 442–451.

McFarland, R., Kirby, D.M., Fowler, K.J., Ohtake, A., Ryan, M.T., Amor, D.J., Fletcher, J.M., Dixon, J.W., Collins, F.A., Turnbull, D.M., et al. (2004a). De novo mutations in the mitochondrial ND3 gene as a cause of infantile mitochondrial encephalopathy and complex I deficiency. Ann. Neurol. *55*, 58–64.

McFarland, R., Taylor, R.W., Elson, J.L., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (2004b). Proving pathogenicity: when evolution is not enough. Am. J. Med. Genet. A. *131*, 107–108; author reply 109–110.

Merkl, R., and Zwick, M. (2008). H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. BMC Bioinformatics *9*, 151.

Mitchell, A.L., Elson, J.L., Howell, N., Taylor, R.W., and Turnbull, D.M. (2006). Sequence variation in mitochondrial complex I genes: mutation or polymorphism? J. Med. Genet. *43*, 175–179.

Montoya, J., López-Gallardo, E., Díez-Sánchez, C., López-Pérez, M.J., and Ruiz-Pesini, E. (2009). 20 years of human mtDNA pathologic point mutations: Carefully reading the pathogenicity criteria. Biochim. Biophys. Acta BBA - Bioenerg. *1787*, 476–483.

Munnich, A., Rötig, A., Chretien, D., Cormier, V., Bourgeron, T., Bonnefont, J.P., Saudubray, J.M., and Rustin, P. (1996). Clinical presentation of mitochondrial disorders in childhood. J. Inherit. Metab. Dis. *19*, 521–527.

Narayanan, A., Keedwell, E.C., and Olsson, B. (2002). Artificial intelligence techniques for bioinformatics. Appl. Bioinformatics *1*, 191–222.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

O., R., Molina-Espiritu, M., Salas, F., Soriano, C., Barrientos, C., S., J., and A., J. (2013). Decoding the Building Blocks of Life from the Perspective of Quantum Information. In Advances in Quantum Mechanics, P. Bracken, ed. (InTech),.

Ohanian, M., Otway, R., and Fatkin, D. (2012). Heuristic Methods for Finding Pathogenic Variants in Gene Coding Sequences. J. Am. Heart Assoc. *1*, e002642.

Pace, C.N., and Scholtz, J.M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. Biophys. J. *75*, 422–427.

Pereira, L., Soares, P., Radivojac, P., Li, B., and Samuels, D.C. (2011). Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. Am. J. Hum. Genet. *88*, 433–439.

Pervez, M.T., Babar, M.E., Nadeem, A., Aslam, M., Awan, A.R., Aslam, N., Hussain, T., Naveed, N., Qadri, S., Waheed, U., et al. (2014). Evaluating the accuracy and efficiency of multiple sequence alignment methods. Evol. Bioinforma. Online *10*, 205–217.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinforma. Oxf. Engl. *18 Suppl 1*, S71–S77.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Saha, I., Maulik, U., Bandyopadhyay, S., and Plewczynski, D. (2012). Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids *43*, 583–594.

Sandler, I., Zigdon, N., Levy, E., and Aharoni, A. (2014). The functional importance of co-evolving residues in proteins. Cell. Mol. Life Sci. CMLS *71*, 673–682.

Schieppati, A., Henter, J.-I., Daina, E., and Aperia, A. (2008). Why rare diseases are an important medical and social issue. Lancet Lond. Engl. *371*, 2039–2041.

Schmidt, T.R., Wu, W., Goodman, M., and Grossman, L.I. (2001). Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. Mol. Biol. Evol. *18*, 563–569.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382–W388.

Simonetti, F.L., Teppa, E., Chernomoretz, A., Nielsen, M., and Marino Buslje, C. (2013). MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res. *41*, W8–W14.

Sluis, E.O. van der, Bauerschmitt, H., Becker, T., Mielke, T., Frauenfeld, J., Berninghausen, O., Neupert, W., Herrmann, J.M., and Beckmann, R. (2015). Parallel structural evolution of mitochondrial ribosomes and OXPHOS complexes. Genome Biol. Evol. evv061.

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M.B. (2009). Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am. J. Hum. Genet. *84*, 740–759.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res. *12*, 1611–1618.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. J. Mol. Biol. *425*, 3919–3936.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med. *1*, 13.

Stone, E.A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. *15*, 978–986.

Thorburn, D.R., and Rahman, S. (1993). Mitochondrial DNA-Associated Leigh Syndrome and NARP. In GeneReviews(®), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J. Bean, T.D. Bird, C.-T. Fong, H.C. Mefford, R.J. Smith, et al., eds. (Seattle (WA): University of Washington, Seattle),.

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. *32*, 358–368.

Tourasse, N.J., and Li, W.H. (2000). Selective constraints, amino acid composition, and the rate of protein evolution. Mol. Biol. Evol. *17*, 656–664.

Tsuji, J., Frith, M.C., Tomii, K., and Horton, P. (2012). Mammalian NUMT insertion is non-random. Nucleic Acids Res. *40*, 9073–9088.

Tuppen, H. a. L., Fattori, F., Carrozzo, R., Zeviani, M., DiMauro, S., Seneca, S., Martindale, J.E., Olpin, S.E., Treacy, E.P., McFarland, R., et al. (2008). Further pitfalls in the diagnosis of mtDNA mutations: homoplasmic mt-tRNA mutations. J. Med. Genet. *45*, 55–61.

Valdar, W.S.J. (2002). Scoring residue conservation. Proteins *48*, 227–241.

Valentino, M.L., Barboni, P., Ghelli, A., Bucchi, L., Rengo, C., Achilli, A., Torroni, A., Lugaresi, A., Lodi, R., Barbiroli, B., et al. (2004). The ND1 gene of complex I is a mutational hot spot for Leber's hereditary optic neuropathy. Ann. Neurol. *56*, 631–641.

Vanhoof, G., Goossens, F., De Meester, I., Hendriks, D., and Scharpé, S. (1995). Proline motifs in peptides and their biological processing. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *9*, 736–744.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics *13 Suppl 4*, S2.

Vithayathil, S.A., Ma, Y., and Kaipparettu, B.A. (2012). Transmitochondrial cybrids: tools for functional studies of mutant mitochondria. Methods Mol. Biol. Clifton NJ *837*, 219–230.

Wallace, D.C. (2007). Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. Annu. Rev. Biochem. *76*, 781–821.

Wang, K., and Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics *7*, 385.

Wang, C.-Y., Kong, Q.-P., Yao, Y.-G., and Zhang, Y.-P. (2006). mtDNA mutation C1494T, haplogroup A, and hearing loss in Chinese. Biochem. Biophys. Res. Commun. *348*, 712–715.

Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A., and Bryant, S.H. (2000). Cn3D: sequence and structure views for Entrez. Trends Biochem. Sci. *25*, 300–302.

Ware, S.M., El-Hassan, N., Kahler, S.G., Zhang, Q., Y-W, Miller, E., Wong, B., Spicer, R.L., Craigen, W.J., Kozel, B.A., et al. (2009). Infantile cardiomyopathy caused by a mutation in the overlapping region of mitochondrial ATPase 6 and 8 genes. J. Med. Genet. *46*, 308–314.

Wasikowski, M., and Chen, X. (2010). Combating the Small Sample Class Imbalance Problem Using Feature Selection. IEEE Trans. Knowl. Data Eng. *22*, 1388–1400.

Worth, C.L., Preissner, R., and Blundell, T.L. (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. *39*, W215–W222.

Yao, Y.-G., Kong, Q.-P., Salas, A., and Bandelt, H.-J. (2008). Pseudomitochondrial genome haunts disease studies. J. Med. Genet. *45*, 769–772.

Yarham, J.W., Al-Dosary, M., Blakely, E.L., Alston, C.L., Taylor, R.W., Elson, J.L., and McFarland, R. (2011). A comparative analysis approach to determining the pathogenicity of mitochondrial tRNA mutations. Hum. Mutat. *32*, 1319–1325.

Yu-Wai-Man, P., and Chinnery, P.F. (1993). Leber Hereditary Optic Neuropathy. In GeneReviews(®), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J. Bean, T.D. Bird, C.-T. Fong, H.C. Mefford, R.J. Smith, et al., eds. (Seattle (WA): University of Washington, Seattle),.

Zarrouk Mahjoub, S., Mehri, S., Ourda, F., Finsterer, J., et al. (2012). Novel m.15434C>A (p.230L>I) Mitochondrial Cytb Gene Missense Mutation Associated with Dilated Cardiomyopathy. Int. Sch. Res. Not*, 2012*, e251723.