



UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL

**Caracterización de las fuerzas evolutivas
que afectan a las proteínas codificadas
por el genoma mitocondrial**

Autor:

D. Héctor Valverde Pareja

Director:

Prof. Dr. Juan Carlos Aledo Ramos

18 de octubre de 2013

*A mis padres.
A mis hermanos.*



Biología Molecular y Bioquímica
Facultad de Ciencias
Universidad de Málaga

Dr. Juan Carlos Aledo Ramos, profesor titular de Biología Molecular y Bioquímica de la Universidad de Málaga,

Certifica:

Que el trabajo de investigación titulado “Caracterización de las fuerzas evolutivas que afectan a las proteínas codificadas por el genoma mitocondrial”, ha sido realizado bajo su dirección en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga por el Licenciado D. Héctor Valverde Pareja, acorde con lo dispuesto en el Capítulo V del Real Decreto 1393/2007.

Málaga, 10 de septiembre de 2013

Prof. Dr. Juan Carlos Aledo Ramos

Este trabajo ha sido subvencionado por el proyecto
CGL2010-18124 perteneciente al Plan Nacional de Investigación
del Ministerio de Ciencia e Innovación.

Índice general

Agradecimientos	vii
Prólogo	ix
Introducción general	1
0.1 Fuerzas evolutivas	1
0.2 ¿Cómo se miden las fuerzas evolutivas?	3
0.3 La mitocondria y la evolución de su genoma	4
0.4 Presión selectiva que afecta a las proteínas de la cadena respiratoria	8
0.5 Evolución de residuos implicados en contactos intermoleculares	9
1 Fuerzas adaptativas sobre el contenido de metionina en mtDNA	11
1.1 Resumen	11
1.2 Introducción	12
1.3 Objetivos	14
1.4 Material y métodos	15
1.4.1 Datos	15
1.4.2 Modelado computacional	15
1.4.3 Tratamiento estadístico	16
1.5 Resultados	17
1.5.1 Existe correlación entre longevidad y frecuencia nucleotídica	17
1.5.2 La composición nucleotídica influye en el contenido de Thr y Met .	17
1.5.3 El contenido de cisteína, metionina y treonina en las proteínas codificadas por mtDNA está sometido a distintas fuerzas evolutivas	18
1.5.4 El análisis estadístico de las diferencias entre las secuencias reales y barajadas revela que el contenido de metionina está sometido a selección positiva	21
1.5.5 Existe una correlación entre la adición de metionina (Δ Met) y la longevidad	21
1.6 Discusión	23
1.6.1 Una nueva aproximación al estudio de la influencia del sesgo mutacional sobre la frecuencia de aminoácidos	24
1.6.2 La cisteína es excluida de las proteínas, pero la variación en su abundancia no depende de la longevidad	25

1.6.3	La variación del contenido de treonina está influenciada exclusivamente por el sesgo mutacional	25
1.6.4	La variación del contenido de metionina está impulsada por mecanismos selectivos además del sesgo mutacional	26
1.6.5	Los animales más longevos incorporan más metioninas en sus proteínas de lo esperado por la composición nucleotídica de sus genes	26
1.6.6	La ausencia de correlación entre longevidad y ΔCys y ΔThr se explica por distintas causas	27
1.6.7	No se observa selección purificadora en contra de la metionina en animales de vida larga, sino una presión adaptativa en animales de vida corta	27
1.6.8	La abundancia de metionina en proteínas puede estar determinada por un compromiso entre su capacidad de eliminar ROS y el efecto nocivo de su oxidación	28
1.7	Conclusiones	29
2	Dinámica evolutiva de citocromo b y COX 1	31
2.1	Resumen	31
2.2	Introducción	31
2.3	Objetivos	33
2.4	Material y métodos	33
2.4.1	Datos y modelado molecular	33
2.4.2	Determinación de las posiciones enterradas y expuestas	34
2.4.3	Determinación de la entropía de Shannon	34
2.4.4	Clasificación de posiciones en base a sus valores de entropía de Shannon	36
2.4.5	Cálculo de tasas de sustituciones	36
2.4.6	Ánáisis de la estabilidad termodinámica	38
2.4.7	Ánáisis estadísticos	38
2.5	Resultados	38
2.5.1	El interior de COX 1 está enriquecido en posiciones invariantes, pero éstas se distribuyen aleatoriamente a lo largo de toda la estructura primaria de citocromo b	40
2.5.2	El interior de COX 1, pero no el de citocromo b, exhibe una entropía de Shannon por debajo de lo justificable por el azar	40
2.5.3	Las posiciones no constreñidas se distribuyen aleatoriamente a lo largo de citocromo b, pero son selectivamente excluidas del interior de COX 1	42
2.5.4	El gen de COX 1 está sujeto tanto a una presión selectiva como a una selección purificadora mayores que las del gen de citocromo b	43
2.5.5	La estabilidad termodinámica puede dar cuenta del comportamiento diferencial de COX 1 y citocromo b	47
2.6	Discusión	49

2.6.1	Existen diferencias en la dinámica evolutiva de citocromo b y COX 1	49
2.6.2	La presión mutacional es mayor sobre COX 1 que sobre citocromo b	50
2.6.3	Las diferencias entre la evolución de citocromo b y COX 1 adquiere mayor relevancia en procesos post-transcripcionales	50
2.6.4	La estabilidad termodinámica está estrechamente relacionada con la dinámica evolutiva de citocromo b y COX 1	52
2.6.5	La influencia de los complejos completos sobre la dinámica evolutiva de sendas subunidades difiere entre citocromo b y COX 1 . . .	53
2.7	Conclusiones	54
3	Evolución de residuos implicados en contactos intermoleculares	55
3.1	Resumen	55
3.2	Introducción	56
3.3	Objetivos	57
3.4	Material y métodos	58
3.4.1	Datos	58
3.4.2	Diseño e implementación del modelo de estudio	58
3.4.3	Caracterización estructural de los residuos	59
3.4.4	Caracterización evolutiva de la subunidades	60
3.4.5	Cambios en la estabilidad termodinámica ($\Delta\Delta G$)	62
3.5	Resultados	62
3.5.1	Comparativa entre los dos modelos de estudio	62
3.5.2	Comportamiento evolutivo de las distintas regiones de las cadenas codificadas por mtDNA	64
3.5.3	Tasas de interacción de subunidades codificadas por mtDNA . . .	65
3.5.4	La estabilidad termodinámica no explica las diferencias entre el comportamiento evolutivo de los residuos en contacto con cadenas mitocondriales y nucleares	66
3.5.5	Los residuos expuestos y libres de COX 1 se encuentran excepcionalmente conservados	69
3.5.6	Existe correlación entre $\Delta\Delta G$ y Σd_N	71
3.5.7	Relevancia de los residuos funcionales de COX 1 en su dinámica evolutiva	72
3.5.8	Tasas de interacción de las cadenas codificadas por nDNA	72
3.5.9	Correlación entre la edad de las proteínas y el valor de d	75
3.6	Discusión	75
3.6.1	Comparación de los dos modelos de cálculo de tasas de interacción	76
3.6.2	¿Por qué los residuos expuestos y libres de COX 1 están tan conservados?	77
3.6.3	Comportamiento diferencial entre interacciones jóvenes y antiguas	78
3.6.4	Comportamiento evolutivo de las subunidades codificadas por nDNA	79
3.7	Conclusiones	81
A	Mecom (<i>Molecular Evolution of protein COMplexes</i>)	83

A.1	Introducción	83
A.2	Implementación	84
A.2.1	Módulo Mecom::Contact	84
A.2.2	Módulo Mecom::Surface	85
A.2.3	Módulo Mecom::Subsets	85
A.2.4	Módulo Mecom::Align::Subset	87
A.2.5	Módulo Mecom::EasyYang	88
A.2.6	Módulo Mecom::Statistics::RatioVariance	88
A.2.7	Módulo Mecom::Report	88
A.2.8	Módulos Mecom y Mecom::Config	88
A.3	Dependencias	89
A.4	Distribución e Instalación	89
A.4.1	Instalación mediante CPAN	90
A.5	Ejemplo de uso	90
A.6	Desarrollo colaborativo	91
B	Modelos de barajado de secuencias	93
C	Análisis de selección positiva en proteínas codificadas por nDNA	95
D	Cálculo de las tasas de sustituciones de mtDNA	97
E	Publicaciones	99
E.1	Publicación del capítulo 1	101
E.2	Publicación del capítulo 2	113
E.3	Publicación del capítulo 3	127
F	Datos	153
F.1	Datos del capítulo 1	153
F.2	Datos del capítulo 2	160
F.3	Datos del capítulo 3	165
F.4	Información estructural	176
F.5	Edad de las proteínas del complejo IV	176
Bibliografía		179

Agradecimientos

«Decía Bernardo de Chartres que somos como enanos a los hombros de gigantes. Podemos ver más, y más lejos que ellos, no por alguna distinción física nuestra, sino porque somos levantados por su gran altura.»

Juan de Salisbury, *Metalogicon* (1159).

El trabajo que se recoge en esta memoria hubiera sido imposible sin la inestimable ayuda de mi director de tesis, Juan Carlos Aledo. Siempre presente y siempre dispuesto a resolver mis dudas en cualquier momento y lugar. Su firmeza, voluntad y perseverancia son, honestamente, admirables. Estas cualidades personales, sumadas a su incuestionable competencia como científico, hacen de él un gran maestro. Gracias, JC.

Mis otros *gigantes* son, afortunadamente, mi familia. Aunque siempre han estado presentes a lo largo de mi vida académica, en esta última etapa he recibido más ánimos por su parte. Han sabido apoyarme justo cuando lo he necesitado y me han hecho saber que comparten conmigo la inminente satisfacción de defender con éxito este trabajo de tesis. Mi hermano David, quizás merece una mención especial por su implicación práctica en el trabajo de investigación. Además de poner a mi entera disposición sus servidores remotos y el asesoramiento que he necesitado (algo patente en el contenido de esta memoria), ha sabido resolverme un sinfín de cuestiones de toda clase, matemática, informática, etc., sin importarle cuánto de su tiempo le he consumido. Muchas gracias, familia.

También quiero agradecer a Jose Ángel Campos y Carolina Cardona, del laboratorio vecino, por ser esos grandes compañeros y amigos del departamento, quienes, desde el primer momento, supieron depositar su confianza en mí sin esperar más a cambio. De igual modo lo hizo Carolina Lobo, quien ha sabido dar los mejores consejos en los mejores momentos. Ya sabes, Carito, que te debo un río con tus peques. Por su parte, Vero e Ian, recientes doctores, han marcado una senda con su ejemplo de seriedad, trabajo y

disciplina. Gracias, pareja, por apoyarme en todo lo que ha estado en vuestra mano. Sinceramente, valoro mucho la atención que he recibido por vuestra parte.

Cristina ha sido la persona que más tiempo ha pasado conmigo. Consecuentemente, ha hecho de sumidero de muchos momentos de estrés. Su actitud y su predisposición a ayudar en todo lo que estuviera en su mano durante la etapa de escritura, ha sido determinante para llevar la tesis a buen fin. Del mismo modo, no puedo dejar de agradecer a sus padres, Carmencita y Manolo, por su desinteresada e inestimable ayuda en nuestro día a día. Gracias de todo corazón.

Aunque la suma de sus edades no llegue a los dedos de una sola mano, quisiera mencionar a Diego y a Saúl. Quizás me di cuenta tarde, pero pasar tiempo con ellos era la mejor forma que encontré de desconectar cuando debía hacerlo. Además, a parte del sentido práctico que, patológicamente intento darle a todo, bien saben los que me rodean que estos niños se han ganado todo mi cariño. Gracias, peques.

Por último, y no menos importante, quisiera mencionar a todos aquellos amigos que han estado pendiente de mi trayectoria y han mostrado un interés sincero. Javi, Leo, Jessi, Vero, Sara, Lucas, Mara y Dani. Gracias, amigos.

Prólogo

El cuerpo de esta tesis doctoral está compuesto por tres trabajos de investigación, cuya motivación común es la de describir las fuerzas evolutivas responsables de la variación en la secuencia de proteínas codificadas por el genoma mitocondrial de mamíferos. El primer trabajo, titulado “*Mutational bias plays an important role in shaping longevity-related amino acid content in mammalian mtDNA-encoded proteins*”, aporta al estudio de la evolución molecular un modelo computacional capaz de discernir entre los efectos de fuerzas evolutivas neutrales y selectivas sobre la abundancia de cada aminoácido en las proteínas estudiadas. Además, se centra en el análisis de la variación de la cantidad de metionina en relación con su potencial función antioxidante. En el segundo, titulado “*Thermodynamic stability explains the differential evolutionary dynamics of Cytochrome b and COX1 in mammals*”, se toman en consideración dos propiedades estructurales: la exposición al solvente y la estabilidad termodinámica; que influyen en el ritmo evolutivo de dos de las proteínas codificadas por el genoma mitocondrial de mamíferos. Por último, el tercer trabajo de investigación, titulado “*Evolution of residues from the Cytochrome c Oxidase complex engaged in intermolecular contacts*”, se añade un nuevo elemento estructural que influye directamente en la dinámica evolutiva de las proteínas: la relación de proximidad entre regiones de la estructura de proteínas que forman parte de un complejo enzimático, teniendo en cuenta, además, el origen genómico de cada subunidad, poniendo así de manifiesto fenómenos de coevolución intergenómica.

Introducción general

Tras la explosión de las técnicas en biología molecular y los talentosos esfuerzos de almacenar y estructurar una ingente cantidad de información biológica, los datos moleculares tomaron protagonismo en los análisis evolutivos. Así, surgió una nueva y eficaz perspectiva sobre el estudio de la evolución, denominada *evolución molecular*. Bajo esta disciplina se desarrolla el cuerpo de esta tesis, cuya meta ulterior consiste en la descripción detallada de las fuerzas que modifican el contenido genético mitocondrial, responsable de una de las funciones más importantes de la célula: la obtención de energía en forma de ATP a partir de otras moléculas reducidas.

0.1 Fuerzas evolutivas

En líneas generales, existen dos fenómenos que dirigen la variación en la secuencia nucleotídica: *la presión mutacional* y *la selección natural*. Aunque existe toda una pléthora de estudios que ponen de manifiesto la presencia de ambos mecanismos evolutivos, existe una gran controversia acerca de cuál de estos dos fenómenos es la verdadera fuerza creadora en el proceso de cambio evolutivo (revisión en Nei, 2005).

La *presión mutacional* es el conjunto de factores que influyen sobre la variabilidad de la secuencia nucleotídica, como los agentes mutagénicos, errores en la replicación del DNA, etc. Cuanto mayor sea la influencia de estos factores sobre la molécula de DNA (sumado a la susceptibilidad de ésta a ser alterada), mayor será la probabilidad de que ocurran mutaciones en el genoma y, en muchas ocasiones, lleguen a transmitirse a la descendencia. Muchos autores afirman que la presión mutacional es la fuerza evolutiva más importante. Según esta teoría, la mayoría de las mutaciones son neutrales, es decir, no confieren ningún incremento o detrimiento en la eficacia biológica¹ de la descendencia, y pueden ser fijados en la población por deriva genética. El genetista Thomas Morgan, en su libro “The Scientific Basis of Evolution” (?), afirmaba literalmente: “Si el nuevo mutante no es ni más aventajado ni menos que el viejo carácter, puede reemplazar o no el viejo carácter, dependiendo en parte del azar; pero si la misma mutación ocurre

¹La expresión «eficacia biológica» es una traducción adaptada del término anglosajón “*fitness*”.

una y otra vez, éste reemplazará, muy probablemente, al carácter original”² (p. 132). No obstante, no fue hasta finales de los 60 cuando la teoría neutralista de la evolución molecular fue formalizada por Kimura (?), justo después del descubrimiento de la entonces inexplicable cantidad de polimorfismos encontrados en proteínas pertenecientes a una misma población.

El otro fenómeno mencionado anteriormente, la *selección natural*, puede ser de dos tipos: purificadora o adaptativa (también negativa o positiva, respectivamente). La selección purificadora es responsable de eliminar los cambios deletéreos, es decir, aquellas mutaciones que provocan una pérdida significativa de la función del gen. La selección adaptativa es el aumento en la eficacia biológica de aquellos cambios que producen una mejora en la función del gen. Los neutralistas no rechazan la presencia de la selección, pero reducen su importancia a un segundo plano (?), constituyendo un “mero tamiz destinado a mantener mutaciones ventajosas y eliminar mutaciones deletéreas” (?).

Por otra parte, las teorías seleccionistas, también englobadas bajo el término “neo-Darwinistas”, sostienen que la selección natural es la fuerza evolutiva más importante en la naturaleza (?; ?; ?; ?). Quizás, los argumentos que cobraron más importancia en la difusión de esta escuela fueron dos. Primero, la mayoría de los genetistas creían que la variabilidad genética contenida en las poblaciones naturales (mayormente obtenida a partir de recombinación) era tan elevada que cualquier cambio puede ocurrir por selección natural sin necesidad de esperar a nuevas mutaciones. Segundo, algunos matemáticos y genetistas habían mostrado que la frecuencia de los cambios genéticos por mutación es mucho menor que el cambio por selección natural.

Desde principios de los 80, el estudio de la evolución estaba cada vez más enfocado en el DNA, lo cual podría esclarecer de una vez por todas cuáles son los mecanismos por los que se rige la evolución, sin embargo, la controversia entre neutralistas y seleccionistas continúa viva. Esto puede estar indicándonos, que el entendimiento de la evolución y la resolución de estos problemas es extremadamente complicado. Sin embargo, es necesario reconocer que muchas de las controversias han sido generadas por ideas erróneas, interpretaciones equivocadas de observaciones empíricas, fallos en análisis estadísticos, etc. (?).

Dado que el objetivo que está presente a lo largo de toda esta tesis es la caracterización de los fenómenos biológicos responsables de la evolución del genoma mitocondrial, nuestro trabajo no puede excluirse de la controvertida dicotomía entre *neutralismo* y *neo-Darwinismo*.

²“If the new mutant is neither more advantageous than the old character, nor less so, it may or may not replace the old character, depending partly on chance; but if the same mutation recurs again and again, it will most probably replace the original character”

0.2 ¿Cómo se miden las fuerzas evolutivas?

Tal y como hemos indicado al principio, haremos hincapié sobre el efecto de la presión mutacional y la selección natural como fuerzas evolutivas responsables de la variabilidad genética. No obstante, existen otros fenómenos responsables de la variación del contenido genético de una población. Tales son, la reproducción sexual, la deriva genética y los movimientos migratorios.

Una de las mayores inquietudes de la biología evolutiva es la detección de procesos adaptativos (selección positiva), ya que estas observaciones pueden ofrecer información sobre otros procesos biológicos relacionados, por ejemplo, con aspectos funcionales. Además, a diferencia de la selección purificadora, los procesos de adaptación están muy acotados en la escala evolutiva, es decir, suceden en intervalos temporales más cortos. Esto es debido a que los procesos de adaptación se dan en respuesta a cambios en el nicho ecológico, mientras que la selección purificadora es responsable de mantener las funciones en ausencia de perturbaciones en el ecosistema.

Los factores primarios implicados en el aumento de la variabilidad genética, son aquellos que producen mutaciones. Éstas, pueden dar lugar, en un determinado gen, a un cambio deletéreo en la proteína para la que codifica y, consecuentemente, ser eliminado de la población. Por el contrario, puede dar lugar a una mejora en la eficacia biológica del organismo, y ser seleccionado en la descendencia. Sin embargo, la mayoría de las mutaciones producen poco o ningún cambio en la eficacia biológica. Del carácter neutral (o casi neutral) de esta última categoría, se deduce que no existe ningún mecanismo selectivo que actúe sobre ellas. Por este motivo, y con el objetivo de excluir la influencia de la selección natural, las diferentes técnicas empleadas en la estimación de la magnitud de la presión mutacional, centran el foco en este tipo de mutaciones. No obstante, no es posible determinar con certeza qué cambios en la secuencia proteica son verdaderamente neutrales. Una forma de evitar esta incertidumbre, es tener en cuenta sólo las mutaciones que no producen cambios en los aminoácidos para los que codifican. Éstas mutaciones se denominan “sustituciones sinónimas” y da una idea cuantitativa de la presión mutacional a la que está sometida dicha secuencia.

Del mismo modo que puede estimarse la cantidad de sustituciones sinónimas S , se puede estimar la cantidad de sustituciones no sinónimas N , es decir, la cantidad de mutaciones que dan lugar a un cambio en el aminoácido para el que codifican. Ambas variables son proporcionales a la longitud de la secuencia, por ello, los parámetros que se usan para comparar la dinámica evolutiva de distintos genes son la *tasa de sustituciones sinónimas por sitio sinónimo*, denotada como d_S , y la *tasa de sustituciones no sinónimas por sitio no sinónimo*, denotada como d_N .

El cociente $\frac{d_N}{d_S}$, también denotado como ω , es un indicador de la presión selectiva a la que está sometida un gen. Un valor valor de $\omega < 1$ indica que, aunque la presión mutacional sea alta, se permiten pocas sustituciones no sinónimas, o dicho con otras palabras, este gen está sometido a selección purificadora. Si, por el contrario, $\omega > 1$, el

valor sugiere que las sustituciones no sinónimas se fijan en la descendencia y que, por tanto, está sometido a selección positiva. Por último, si $\omega = 1$, no existe presión selectiva sobre el gen, es decir, los cambios son neutrales. No obstante, es necesario proceder con cautela al interpretar estos resultados. Aunque estas aproximaciones a la caracterización evolutiva de un gen son muy útiles y están muy bien implementadas en multitud de software, estos parámetros tienen algunas limitaciones. Un ejemplo es en el caso de la evolución de las regiones no codificantes que, debido a que no dan lugar a proteínas, las sustituciones no sinónimas no están sometidas a una presión selectiva comparable a la de regiones codificantes. Otro caso es la disparidad en el comportamiento evolutivo entre distintas regiones de un mismo gen, por ejemplo, un valor de $\omega = 1$ puede ser interpretado como una relajación de la presión selectiva o bien como una quimera de selección positiva y purificadora en el mismo locus. Para solventar, en la medida de lo posible, estos inconvenientes, se han desarrollado otras aproximaciones estadísticas destinadas a la detección de procesos adaptativos (?), entre las que se incluye la determinación de ω para codones aislados.

La elección de las herramientas computacionales y estadísticas destinadas a describir la dinámica evolutiva de una muestra, depende en gran medida de las características de ésta. Los objetivos de esta tesis están relacionados con la evolución del genoma mitocondrial de mamíferos. Afortunadamente, hasta hoy se han secuenciado varios centenares de estos genomas, por lo que el tamaño muestral es favorable. Además, la casi total ausencia de regiones no codificantes facilita los procesos de análisis.

0.3 La mitocondria y la evolución de su genoma

La mitocondria es un orgánulo ubicuo en los sistemas eucariotas. Aunque también está implicada en procesos de apoptosis, envejecimiento, transducción de señales y en la proliferación celular (?; ?; ?; ?), su función principal es la producción aeróbica de energía en forma de ATP a través de la fosforilación oxidativa (OXPHOS).

Un subconjunto de las proteínas que forman parte de la cadena respiratoria, está codificado en el genoma mitocondrial. Este genoma es, dado su origen endosimbiótico, similar a los genomas bacterianos, compuesto por una sola molécula circular bicanalicular. Su arquitectura varía entre los distintos reinos de eucariotas. Mientras que en plantas tiene un tamaño de 180 a 600 kb, en animales presenta un rango de 14 a 20 kb. Además, la fracción de DNA no codificante también es muy baja en animales (0.05 - 0.10) en comparación con plantas (0.72), llegando a casi cero en el caso de mamíferos, donde no existen ni intrones ni regiones intergénicas (?). Cada una de las dos hebras tiene una densidad distinta al ser centrifugadas en un gradiente de CsCl (?), y han recibido una denominación referente a este hecho: cadenas ligera y pesada o cadenas L y H, respectivamente. En vertebrados, es muy estable en cuanto a su carga genética. Tiene una longitud de unas 16.5 megabases y aproximadamente el 95 % de su secuencia es codificante. El resto, una pequeña región denominada bucle D ó *D-loop*, contiene el origen de replicación de la

0.3. La mitocondria y la evolución de su genoma

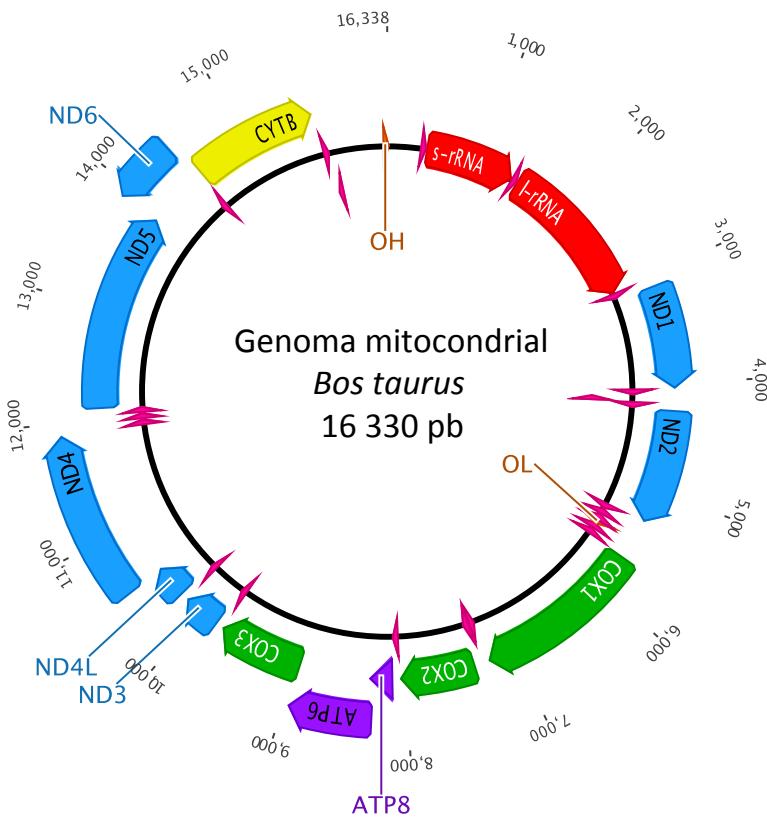


Figura 1: Genoma mitocondrial de *Bos taurus*. Los genes que codifican para los complejos responsables de la fosforilación oxidativa, se representan mediante flechas de distintos colores: en azul, los 7 genes que codifican para el complejo I; en amarillo, citocromo b, el único gen codificado por mtDNA que pertenece al complejo III; en verde, tres subunidades del complejo IV (COX 1, 2 y 3); y en violeta dos genes del complejo V. Adicionalmente se encuentran, en rojo, los dos genes que codifican para el ribosoma mitocondrial, en rosa los 22 tRNAs y en marrón, los dos orígenes de replicación. Las flechas dextrógiros representan los genes codificados en la cadena H y las flechas levógiros representan los de la cadena L.

hebra pesada (O_H) y elementos reguladores de la expresión génica. El genoma completo contiene un total 37 genes que codifican para 13 proteínas implicadas en la fosforilación oxidativa (véase figura 1), dos rRNAs correspondientes a las subunidades 12S y 16S de los ribosomas mitocondriales y 22 tRNAs. La mayoría se encuentran codificados en la cadena H, salvo 8 tRNAs y 1 polipéptido (el gen ND6).

El proceso de replicación del genoma mitocondrial es algo peculiar. La síntesis del nuevo material genético comienza en un origen distinto para cada hebra: O_H en el caso de la hebra pesada y O_L en el de la hebra ligera. El proceso comienza con la replicación de parte de la hebra pesada, y una vez que O_L queda expuesto, comienza la replicación de la hebra ligera. Se dice que este tipo de replicación es asimétrica. La hebra pesada permanece durante más tiempo que la hebra ligera en estado de monohebra durante la

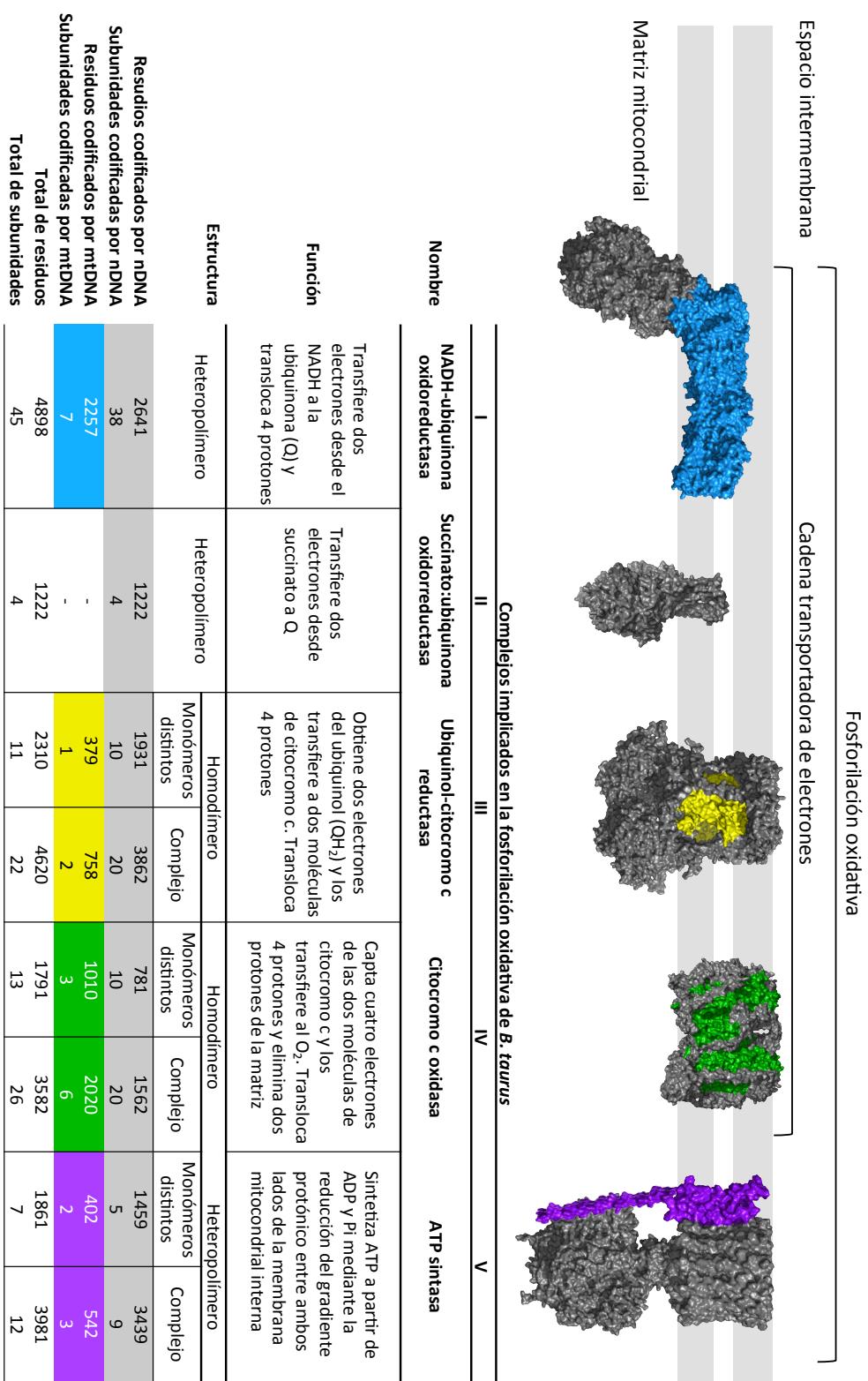


Figura 2: Esquema-resumen de la estructura de la fosforilación oxidativa.

0.3. La mitocondria y la evolución de su genoma

replicación (?).

La cadena respiratoria está formada principalmente por 4 complejos proteicos (figura 2). Brevemente, el proceso de respiración celular consiste en una transferencia indirecta de electrones desde compuestos reducidos, provenientes del ciclo de Krebs, hasta el oxígeno. El paso de esta corriente eléctrica a través de los distintos elementos de la cadena respiratoria, produce un bombeo de protones en contra de gradiente hacia el espacio intermembrana, lo que supone un aumento del potencial protónico, el cual, es posteriormente disipado por la ATP sintasa para formar ATP a partir de ADP y Pi, constituyendo entonces el mencionado proceso de fosforilación oxidativa. Cada uno de estos cinco complejos está formado por varios polipéptidos codificados por dos genomas distintos, el nuclear (nDNA) y el mitocondrial (mtDNA). El complejo I o NADH-ubiquinona oxidoreductasa contiene 7 de las 13 proteínas codificadas por mtDNA (ND1, ND2, ND3, ND4, ND4L, ND5, ND6): ND2, ND4 y ND5 están implicadas en el transporte electrónico, mientras que ND1 y ND2 juegan un papel estructural importante en la inclusión del complejo en la membrana (?). El complejo II o succinato:ubiquinona oxidoreductasa está compuesto enteramente por subunidades codificadas por el genoma nuclear. El complejo III o ubiquinol-citocromo c reductasa contiene una sola subunidad derivada de mtDNA, citocromo b (cytb), y cumple una función catalítica esencial: la reducción de citocromo c. En el complejo IV o citocromo c oxidasa, COX 1 cataliza la transferencia de electrones al último acceptor, el oxígeno molecular; COX 2 y COX 3 también pertenecen al núcleo catalítico. Estas tres son las únicas subunidades de este complejo codificadas en el genoma mitocondrial. Por último, el complejo V (ATP sintasa) contiene dos proteínas codificadas en la mitocondria: ATP6, un componente clave en el canal protónico (F_0) y ATP8 que parece ser un regulador del ensamblaje del complejo (?). El resto de las subunidades que constituyen la fosforilación oxidativa están codificadas en el genoma nuclear (nDNA), son sintetizadas en el citoplasma, y posteriormente transportadas a la mitocondria.

La importancia de estas proteínas en el funcionamiento global de la célula, debido a que constituyen su principal fuente energética, es bien conocida. Por este motivo, se esperaba que la evolución del proteoma mitocondrial presentara una variabilidad muy reducida. Sin embargo, en 1979 se observó que, en realidad, el genoma mitocondrial evoluciona incluso más deprisa que el genoma nuclear (?), sobre todo en el caso de las mutaciones sinónimas. Esta alta presión mutacional al que está sometido el mtDNA, ha sido asociada con la presencia de un sistema de reparación deficiente (?), la ausencia de proteínas similares a las histonas y el peculiar modelo de replicación mitocondrial, en el que la hebra H se encuentra en estado de monohebra durante mayor tiempo, expuesta a daño hidrolítico y oxidativo. En consecuencia, es más proclive a sufrir mutaciones (?). En líneas generales, se considera que el genoma mitocondrial es una molécula susceptible al cambio, tanto por su estructura como por su mecanismo de síntesis.

Atendiendo a la composición nucleotídica del mtDNA, las proporciones de las bases se alejan de lo previsible según la segunda ley de paridad de Chargaff (PR2), que indica que, en una misma hebra debe haber una proporción semejante entre bases púricas

y pirimidínicas (ej. %A %T y %G %C). La violación de la PR2 se denomina *sesgo nucleotídico*, y da lugar a un cambio en la frecuencia de codones. En el caso del genoma mitocondrial, la hebra pesada presenta un bajo contenido en G, especialmente atendiendo a posiciones sujetas a baja selección natural (tripletes degenerados), mientras que se observa una alta proporción de A, por tanto, los codones más frecuentes son aquellos que contienen A y no G. Este desequilibrio provoca, por consiguiente, un cambio en la frecuencia aminoacídica en las proteínas codificadas en el genoma mitocondrial. De hecho, existen numerosos trabajos en el que se establece que la evolución de proteínas codificadas por mtDNA está afectada por la composición nucleotídica de sus genes (?; ?; ?; ?; ?). Reyes y colaboradores (?) hallaron una correlación entre el sesgo nucleotídico y el tiempo que permanece expuesto el DNA de mitocondrias de distintas especies de mamíferos, por lo que se aceptó el particular mecanismo de replicación del mtDNA como una explicación razonable de la variación del contenido GC.

Aunque es cierto que, en comparación con el nDNA, el mtDNA presenta una mayor presión mutacional, atendiendo al cambio en forma de mutaciones no sinónimas, es decir, mutaciones que provocan una sustitución aminoacídica en el péptido para el que codifica, las tasas de cambio son comparables. Esto quiere decir, que aunque el genoma mitocondrial esté sometido a una elevada presión mutacional, también existen mecanismos de selección purificadora responsables del mantenimiento de la función de las proteínas codificadas (?).

0.4 Presión selectiva que afecta a las proteínas de la cadena respiratoria

El ritmo al que evolucionan las proteínas difiere en gran medida entre unas y otras (?). La tasa a la que una proteína acumula sustituciones depende de la presión mutacional, de sus propiedades biofísicas y bioquímicas, y de la intensidad de la selección que actúe sobre su función. No obstante, recientemente se ha demostrado que uno de los mayores determinantes de la evolución de una proteína es, sorprendentemente, su nivel de expresión (?; ?). Tras este descubrimiento, numerosos grupos han investigado cuáles son las causas que pueden explicar la relación entre la tasa evolutiva de una proteína y la concentración de mRNA (?; ?; ?; ?), dando lugar a distintas teorías no excluyentes entre sí. Estas teorías, las cuales se discuten más adelante (página 50), desplazan la importancia de la función de la proteína, como determinante de su capacidad evolutiva, a un segundo plano. A cambio, estos autores sugieren que la eficiencia de plegado (?), el coste energético de la expresión (?; ?) y las interacciones inespecíficas (?), influyen en mayor medida sobre la tasa evolutiva de las proteínas.

Otro factor a tener en cuenta, en el caso de las proteínas mitocondriales, es el entorno oxidativo en el que se encuentran. En la mitocondria, se produce gran cantidad de especies reactivas de oxígeno (ROS), por lo que estas proteínas deben estar sometidas a una presión selectiva por estrés oxidativo. De hecho, existen numerosos estudios que

0.5. Evolución de residuos implicados en contactos intermoleculares

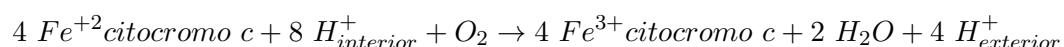
ponen de manifiesto distintas estrategias que han surgido ante la potencial necesidad de proteger el proteoma mitocondrial (?; ?; ?; ?). El trabajo desarrollado en el capítulo 1 se centra en el rol de los aminoácidos más susceptibles de ser oxidados, metionina y cisteína, en la resistencia a estrés oxidativo. En él, el objetivo principal es detectar si existe variabilidad en la abundancia de metionina en proteínas codificadas por mtDNA, en relación con la intensidad de la presión selectiva provocada por el estrés oxidativo.

En este escenario, no es de extrañar que los residuos expuestos al solvente de las proteínas mitocondriales, estén sometidos a distintas restricciones evolutivas que los residuos del interior, bien por su exposición a daño oxidativo mediado por ROS, o bien por otros motivos estructurales o funcionales. Aunque este hecho no se ha demostrado hasta la fecha en proteínas mitocondriales, sí se han llevado a cabo estudios en levaduras (?; ?; ?) y bacterias (?) que apuntan a una evolución diferencial entre superficie e interior. No obstante, estas diferencias parecen estar relacionadas con la contribución a la estabilidad termodinámica del conjunto (?). En el capítulo 2, se estudian las diferencias en la dinámica evolutiva entre la superficie y el interior de dos proteínas mitocondriales (citocromo b y COX 1) en relación con la contribución a la estabilidad termodinámica de cada residuo.

0.5 Evolución de residuos implicados en contactos intermoleculares

Remitiéndonos de nuevo a la figura 2, se ilustra mediante colores cómo los complejos I, III, IV y V de la fosforilación oxidativa de mamíferos están formados por proteínas codificadas por dos genomas distintos, el genoma nuclear (nDNA) y el genoma mitocondrial (mtDNA). Es esperable, entonces, que las interacciones entre residuos codificados por genomas distintos tenga una repercusión sobre la dinámica evolutiva de éstos, en comparación con otros aminoácidos que no están sometidos a estas restricciones evolutivas. De hecho, trabajos previos han puesto de manifiesto la importancia de la adaptación intergenómica (?; ?; ?). Algunos autores han llevado a cabo trabajos cuyo objetivo fue el de describir los mecanismos evolutivos predominantes en las regiones de contacto del complejo IV (?; ?; ?), puesto que es un modelo de estudio muy adecuado gracias a la disponibilidad de datos estructurales y genéticos.

El complejo citocromo c oxidasa está formado por dos subunidades iguales. Éstas, a su vez, están formadas por 13 proteínas, 10 codificadas por nDNA y 3 por mtDNA. Además, el complejo posee dos grupos hemos, un citocromo a y otro a₃, así como dos centros de cobre (uno denominado *centro Cu_A* y el otro *centro Cu_B*). La reacción catalizada por el complejo puede resumirse como sigue:



Los subíndices *interior* y *exterior* hacen referencia a la localización de los protones, dependiendo de si se encuentran en la matriz mitocondrial o en el espacio intermembrana, respectivamente. Esta reacción se produce en lo que se conoce como el *núcleo catalítico* del enzima que, en este caso, está formado por las tres subunidades codificadas por mtDNA: COX 1, COX 2 y COX 3. La función del resto de las cadenas que forman el complejo es, generalmente, desconocida, aunque existen algunos trabajos que sugieren que estas proteínas cumplen una función de protección del núcleo catalítico ante el potencial efecto desestabilizador del solvente³ (?). Así pues, en el complejo citocromo c oxidasa, podemos encontrar aminoácidos sometidos a restricciones evolutivas muy dispares, ya sea por su implicación en la función catalítica, sus propiedades estructurales o su origen genómico. En un esfuerzo de describir la dinámica evolutiva de distintas regiones del complejo IV, en el capítulo 3, hemos aplicado distintos criterios de clasificación de residuos y, posteriormente, hemos analizado cada región de forma independiente.

³La expresión «potencial efecto desestabilizador del solvente» es una traducción adaptada de «fold-disruptive hydration».

Capítulo 1

Existen fuerzas selectivas que alteran la abundancia de metionina en proteínas codificadas por el genoma mitocondrial de mamíferos

1.1 Resumen

Durante la evolución, cambios en la composición aminoacídica han podido dar lugar a proteomas mitocondriales mejor dotados frente a estrés oxidativo. Sin embargo, debido al difícil problema de distinguir entre restricciones o adaptaciones funcionales de las secuencias proteicas y los cambios en éstas provocados por el sesgo mutacional, el valor adaptativo de la variación aminoacídica continúa bajo discusión. En este trabajo hemos analizado las secuencias de mtDNA codificantes de 173 especies de mamíferos, aislando el efecto de la composición nucleotídica sobre la abundancia de aminoácidos en las proteínas. Tal y como se esperaba, encontramos una notable exclusión de residuos cisteínil, aunque no se apreciaba ninguna relación entre el contenido de cisteína y la longevidad. Por otro lado, la abundancia de treonina observada en las proteínas codificadas en el genoma mitocondrial, está completamente influenciada por la composición nucleotídica. Por último, con respecto a la frecuencia de metionina en estas proteínas, nuestros resultados sugieren que, además de la influencia del sesgo mutacional, existen fuerzas selectivas que modifican el contenido de residuos metionil. No obstante, el rol del estrés oxidativo como fuerza selectiva sobre la abundancia de metionina continúa en debate.

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

1.2 Introducción

Según la teoría de los radicales libres, el daño causado por especies reactivas de oxígeno (ROS) sobre macromoléculas acaba en un declive funcional asociado con el envejecimiento en mamíferos (?). De acuerdo con esta teoría, la mitocondria juega un papel fundamental en el envejecimiento debido a que es, al mismo tiempo, una fuente de ROS (a causa del funcionamiento de la cadena transportadora de electrones) y una diana del daño oxidativo, lo que podría dar lugar a una reducción en sus funciones metabólicas (?). En la misma línea de esta premisa, se han descrito en la literatura, diversas adaptaciones mitocondriales al estrés oxidativo (?; ?; ?).

Una primera línea defensiva frente al daño oxidativo, consiste en aminorar la tasa de producción de ROS en la mitocondria (?). En este sentido, estudios comparativos entre especies han puesto de manifiesto la implicación del complejo I de la cadena transportadora de electrones (?; ?; ?) y proteínas desacoplantes (?; ?) como dos dianas potenciales de selección natural. Otra estrategia que parece haberse adoptado en animales durante el curso de la evolución, consiste en hacer las macromoléculas menos susceptibles a oxidación. Este hecho ha sido bien ilustrado por Pamplona y colaboradores (?), quienes describieron una relación inversa entre la susceptibilidad de los lípidos de membrana a peroxidación y la longevidad en mamíferos. Más recientemente, este mismo grupo ha descrito también una correlación negativa entre longevidad y el potencial de oxidación de ácidos grasos libres en el plasma de diversas especies de mamíferos (manuscrito bajo revisión). En la misma línea, diferentes autores sugieren que los cambios ocurridos en el genoma mitocondrial están dirigidos por la presencia de una presión selectiva, hacia un proteoma más preparado para resistir el estrés oxidativo (?; ?; ?).

Aunque todos los aminoácidos proteinogenéticos son dianas potenciales de daño oxidativo, aquellos que contienen grupos sulfurados (cisteína y metionina) son particularmente sensibles a oxidación (?). Un punto interesante a este respecto, es que tanto para la cisteína como para la metionina, se ha descrito una correlación negativa entre su abundancia y la longevidad (?; ?). Sin embargo, mientras que a la cisteína se le atribuye un rol pro-oxidante, diversos autores consideran que la metionina juega un papel anti-oxidante en el proteoma mitocondrial, tal y como argumentan estudios precedentes a esta tesis llevados a cabo por el mismo grupo de investigación (?). La causa por la que la abundancia de cisteína va en detrimento de la resistencia a oxidación por parte de la proteína en la que se encuentra, se halla en su potencial de formar enlaces intra- e interproteicos, llevando de forma irreversible a la inactivación y posterior degradación de la proteína. Este hecho es causa de la selección purificadora en contra de los residuos cisteinil. Por el contrario, el principal producto de la oxidación de la metionina es la metionina sulfóxido (MetO), la cual puede volver a ser reducida por la metionina sulfóxido reductasa, con el consumo de una molécula de NADPH (?). De esta forma, un equivalente de ROS es eliminado del medio por cada residuo de metionina reparado (figura 1.1; Levine et al., 1996). Como tal, un enriquecimiento en metionina en el proteoma mitocondrial podría representar una respuesta adaptativa a estrés oxidativo (?;

?). Se sabe que los animales de vida corta están sometidos a un mayor estrés oxidativo y, por tanto, existe una mayor presión selectiva que favorece el desarrollo de mecanismos de eliminación de ROS. En base a esta idea, los animales de vida corta deberían estar sujetos a una mayor presión selectiva que incremente el contenido de metionina en las proteínas mitocondriales. En otras palabras, si los residuos de metionina sirven como sumidero de ROS en la mitocondria, entonces las proteínas de animales sujetos a un mayor estrés oxidativo (animales de vida corta) deberían acumular metionina de forma más activa que sus ortólogos en especies expuestas a un menor estrés oxidativo. Esta hipótesis es reminiscente de la relación negativa entre niveles de antioxidantes endógenos y longevidad (revisado en Pamplona y Costantini, 2011).

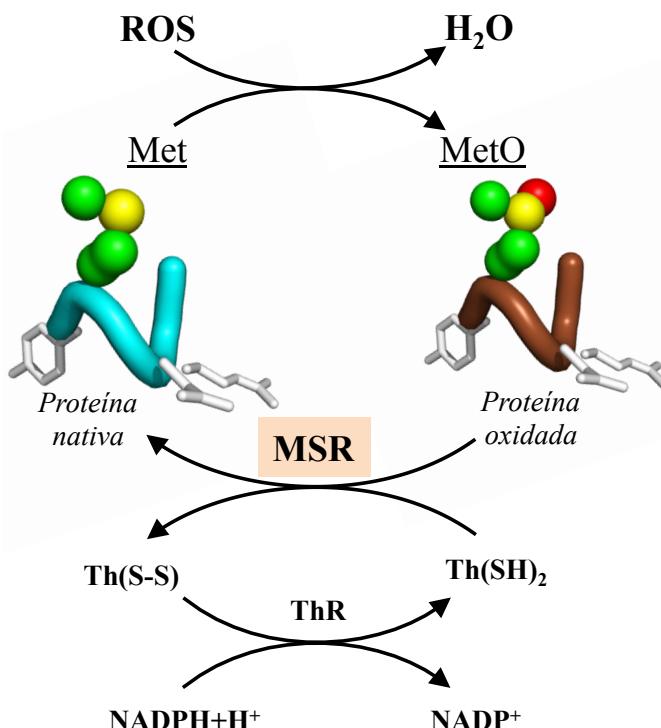


Figura 1.1: Mecanismo propuesto por Levine y colaboradores (?) por el que se le atribuye un rol antioxidante a la metionina. La proteína nativa (izquierda) contiene una metionina susceptible (Met) que es oxidada a metionina sulfóxido (MetO) por ROS, produciendo, en ocasiones, un cambio conformacional en la proteína que puede derivar en un declive funcional (derecha). La acción del enzima metionina sulfóxido reducasa (MSR), reduce la MetO devolviendo a la proteína a su conformación nativa, transfiriendo los electrones desde la tiorredoxina reducida ($\text{Thr}(\text{SH})_2$). Posteriormente, la tiorredoxina reducasa (ThR), reduce la tiorredoxina oxidada ($\text{Thr}(\text{S-S})$) a $\text{Thr}(\text{SH})_2$, consumiendo una molécula de poder reductor (NADPH).

Por otra parte, se ha observado una correlación positiva entre la abundancia de treonina en regiones transmembrana y longevidad (?; ?). Ya que la treonina provee a las hélices de una mayor cantidad de enlaces de hidrógeno intracatenarios, incrementando la estabilidad de la proteína, se ha sugerido que un incremento en la abundancia de este

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

residuo podría ser beneficioso para lograr una mayor longevidad (?).

Aunque estas correlaciones entre la longevidad y abundancia de residuos (negativa en el caso de cisteína y metionina, y positiva en el caso de la treonina) fueron originalmente interpretadas en términos de selección Darwinista, Jobson y colaboradores (?) arrojaron dudas sobre el carácter adaptativo de los cambios en la frecuencia aminoacídica explicados anteriormente. Estos autores explican estos cambios en la composición de la secuencia proteica mediante procesos neutrales, en concreto, cambios en la composición nucleotídica y sesgos en la frecuencia de uso de los distintos codones, que tiende a ser muy relevante en genomas mitocondriales (?). En línea con esta objeción, numerosos trabajos demuestran que la evolución de proteínas está afectada por la composición nucleotídica de sus correspondientes genes (?; ?; ?; ?; ?), y el genoma mitocondrial no es una excepción a esta regla (?; ?; ?; ?). De hecho, se ha demostrado que el sesgo mutacional puede mimetizar los efectos de la selección positiva (?). Estos precedentes ponen de manifiesto las dificultades que existen para distinguir entre las restricciones funcionales sobre la secuencia proteica, y la influencia de la composición nucleotídica sobre las mismas. Por esta razón, es necesario proceder con extrema cautela cuando se comparan resultados procedentes de grupos de especies cuyos genomas presentan frecuencias nucleotídicas distintas. Además, se pone de manifiesto la necesidad de nuevas técnicas que permitan discriminar entre fuerzas evolutivas selectivas y neutrales.

En este trabajo, hemos abordado el potencial efecto de la selección natural, a través de la presión selectiva ejercida por estrés oxidativo, sobre el contenido de aminoácidos en proteomas mitocondriales. Para este propósito, hemos desarrollado un modelo de estudio capaz de diseccionar los efectos neutrales del sesgo mutacional de los efectos provenientes de fenómenos adaptativos.

1.3 Objetivos

Ante el escenario que se ha planteado anteriormente, podemos resumir los objetivos de este capítulo en el siguiente orden:

1. Construir y contrastar un modelo válido para el análisis de la influencia del sesgo mutacional en el contenido aminoacídico de proteínas.
2. Aislar el efecto del sesgo mutacional de las fuerzas selectivas que contribuyan a alterar el contenido de metionina en las proteínas llevadas a estudio.
3. Observar la tendencia a incorporar o sustraer residuos metionil de las proteínas estudiadas en relación con la longevidad y, por consiguiente, con el estrés oxidativo del organismo en el que se encuentran codificadas.

1.4 Material y métodos

1.4.1 Datos

Todas las secuencias analizadas en este estudio fueron obtenidas del NCBI (*National Center for Biotechnology Information*). De la base de datos *genome*, se recolectaron los genomas mitocondriales de 173 especies de mamíferos. Además, para cada especie, se obtuvo el valor de máxima longevidad (*MLSP*) de la base de datos *AnAge* en <http://genomics.senescence.info/species/> (?). En el apéndice F se encuentra un catálogo detallado de los identificadores de cada genoma así como la información obtenida de *AnAge*.

1.4.2 Modelado computacional

Para los análisis de las fuerzas selectivas en el contenido de residuos sulfurados en el genoma mitocondrial, los 12 genes de la hebra ligera fueron concatenados en una sola secuencia nucleotídica lineal. Previamente, se eliminaron todos los codones de parada y algunos más de la región terminal de genes de algunas especies, de forma que la longitud de cada secuencia fuera igual a la del ortólogo más corto. Así, la muestra original se componía de 173 secuencias ortólogas de la misma longitud, 8796 nucleótidos, que codifican para una hipotética proteína de 2932 aminoácidos. A partir del barajado de cada una de las secuencias, se generaron seis modelos distintos:

1. Modelo homogéneo: se barajaron todos los nucleótidos sin generar codones de stop. El resultado fueron secuencias aleatorias con la misma frecuencia nucleotídica que las originales.
2. Modelo “1-2-3”: Se barajaron todos los nucleótidos de forma que cada posición de los tripletes mantuviera la misma frecuencia nucleotídica que la posición equivalente en su secuencia original.
3. Modelos “1”, “2” y “3”: Sólo se barajaron los nucleótidos de la primera, la segunda o la tercera posición de cada triplete, respectivamente.
4. Modelo “1-3”: Se barajaron los nucleótidos de la primera y la tercera posición manteniendo la misma frecuencia nucleotídica que su secuencia original.

En la figura 1.2 se esquematiza un ejemplo del modelo homogéneo, que ilustra el método que se ha seguido para generar las secuencias aleatorias. Paralelamente, y dado que uno de los aminoácidos de más interés del experimento fue la metionina, se construyeron unos conjuntos idénticos de secuencias excluyendo el codón de inicio de la transcripción de cada gen, que en la mayoría de los animales codifica para este aminoácido. Cada una de las secuencias de estos modelos codifica para un número determinado de cada aminoácido. Sobre estos recuentos se realizaron los análisis estadísticos pertinentes para contrastar la validez de nuestras hipótesis.

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

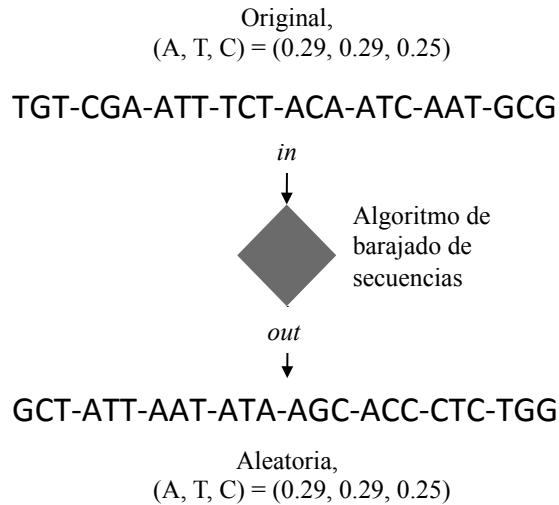


Figura 1.2: Esquema del modelo homogéneo de barajado de secuencias. A partir de una secuencia (original) se cambian las posiciones de los nucleótidos de forma iterativa evitando la aparición de codones de parada. Así, se obtiene otra secuencia aleatoria con la misma frecuencia nucleotídica (representada entre paréntesis) que la original. Con este método, la secuencia queda liberada de cualquier proceso selectivo y la abundancia de los aminoácidos para los que codifique dependerá únicamente de la composición nucleotídica.

1.4.3 Tratamiento estadístico

Para cada especie (i), se computó el número de veces que se encuentra codificado cada aminoácido (metionina, cisteína o treonina) en la secuencia original (x_i) y aleatoria (y_i). Así, para cada aminoácido se obtuvieron n pares ordenados de la forma (x_i, y_i) , siendo n el número de especies analizadas. En ausencia de restricciones funcionales sobre las secuencias proteicas, la hipótesis nula afirma que x_i e y_i tienen la misma probabilidad de ser una mayor que la otra ($P[X < Y] = P[X > Y] = 0.5$). Los pares cuyas coordenadas eran iguales se omitieron. Así, el número de pares totales fue de $n = 173$ para cisteína, $n = 171$ para metionina y $n = 167$ para treonina. Entonces, se calculó la variable aleatoria W como el número de ocurrencias de $x_i - y_i > 0$. Bajo las condiciones de la hipótesis nula, W sigue una distribución binomial, $W \sim Bi(n, 0.5)$. Sin embargo, ya que los valores de n son altos se puede aproximar a una distribución normal. Para ello, se tipificó la variable W teniendo en cuenta que la media es $n/2$ y la varianza es $n/4$ ($E[W] = np = n/2$; $Var[W] = np(1-p) = n/4$). La variable tipificada, Z , sigue entonces una distribución normal, $Z \sim N(0, 1)$.

Los cálculos de probabilidad se realizaron con el programa Wolfram Mathematica 8.0. El resto de los análisis estadísticos se llevaron a cabo con SPSS 15.0.

1.5 Resultados

Con el ya mencionado objetivo de esclarecer las fuerzas evolutivas que modifican el contenido de Cys, Thr y Met en proteínas codificadas por el genoma mitocondrial, se llevó a cabo un análisis compuesto por una serie de simulaciones bioinformáticas de manipulación de secuencias. En primer lugar se estableció la relación entre la longevidad de la especie y la frecuencia nucleotídica de su correspondiente secuencia, y posteriormente se observó que el sesgo mutacional influye en el contenido de Met y Thr en las proteínas codificadas por mtDNA. No obstante, con la idea de que pudieran existir otras fuerzas responsables de la variación en la abundancia de estos residuos, se llevaron a cabo otros análisis en los que se advirtió la presencia de una presión positiva sobre el contenido de Met, mientras que la cantidad de Thr varía conforme cambia la frecuencia nucleotídica. Estos resultados fueron contrastados estadísticamente mediante diversos análisis.

1.5.1 Existe correlación entre longevidad y frecuencia nucleotídica

En la figura 1.3 se representan las relaciones existentes entre la longevidad de las 173 especies de mamíferos analizadas y la cantidad de cada uno de los nucleótidos en los 12 genes de la hebra ligera del genoma mitocondrial. Se observaron correlaciones negativas fuertemente significativas en los casos de la timina y la citosina ($p = 7 \times 10^{-10}$ y 10^{-12} , respectivamente), muy distantes de los casos de la adenina y la guanina ($p = 0.001$ y 0.070 , respectivamente). Unos resultados similares fueron publicados por Samuels (?) en el contexto de la susceptibilidad diferencial del mtDNA al deterioro, aunque con un menor número de secuencias analizadas.

1.5.2 La composición nucleotídica influye en el contenido de Thr y Met

Se calculó la correlación entre la cantidad de residuos codificados por las secuencias aleatorias y la longevidad de la especie. Los resultados para los aminoácidos Cys, Thr y Met se muestran en la figura 1.4. Se obtuvieron valores de correlación significativos para Met y Thr ($r = -0.496$, $p = 4 \times 10^{-12}$ y $r = 0.430$, $p = 3 \times 10^{-9}$, respectivamente) pero no para Cys ($r = -0.040$, $p = 0.598$), independientemente del modelo usado para barajar las secuencias (véase apéndice B). Dado que existe una correlación entre la composición nucleotídica y la longevidad, se deduce que la abundancia de estos aminoácidos viene dada, al menos en parte, por la composición nucleotídica del genoma. Estos resultados son coherentes con los presentados previamente por Jobson y colaboradores (?), con los que sugieren que el sesgo mutacional, o una selección purificadora poco eficiente, es el principal fenómeno responsable de la variación en la abundancia de Cys, Thr y Met en proteínas codificadas por mtDNA. Sin embargo, surge la duda de si podrían existir otras fuerzas que contribuyan a la variación del uso de estos residuos.

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

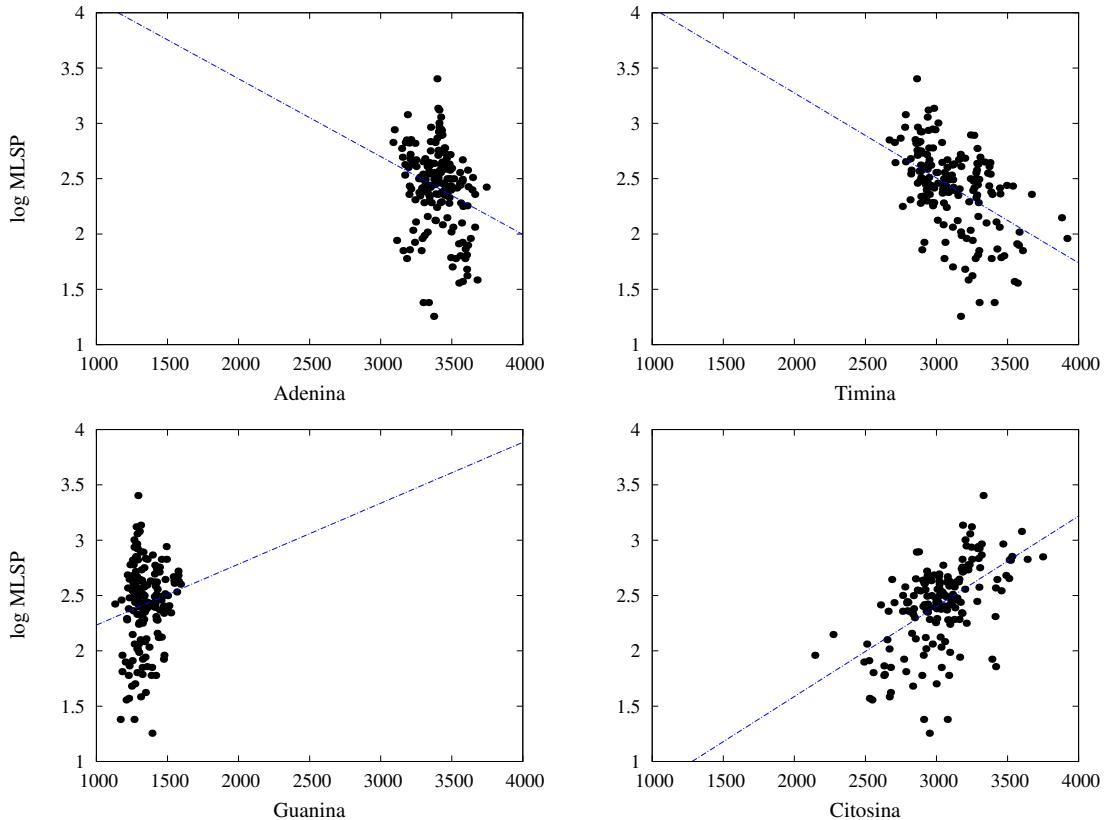


Figura 1.3: Relación entre la abundancia de cada nucleótido en la hebra ligera de mtDNA y la longevidad. Se computó el número de veces que aparece cada nucleótido en los 12 genes codificantes de la hebra ligera del genoma mitocondrial de cada una de las 173 especies. Posteriormente, se estudió la correlación entre la abundancia nucleotídica y *log MLSP*. La abundancia de timina, al igual que la de citosina, mostraron una alta correlación con la longevidad (*p* valores: 7×10^{-15} y 10^{-12} , respectivamente), mientras que adenina y guanina mostraron poca o ninguna (*p* valores: 0.001 y 0.070, respectivamente).

1.5.3 El contenido de cisteína, metionina y treonina en las proteínas codificadas por mtDNA está sometido a distintas fuerzas evolutivas

En una primera aproximación a la duda arrojada en el apartado anterior, se computó el número de veces que aparece cada uno de los residuos en las 12 proteínas codificadas en la hebra ligera, y se representó en tres gráficas de dispersión (una para cada aminoácido) frente al número de veces que aparece codificado en una secuencia aleatoria con la misma frecuencia nucleotídica que la original (figura 1.5). La distribución de los puntos a lo largo del eje de abscisas, correspondiente a la cadena aleatoria, se da acorde con la probabilidad de encontrar un codón codificante para dicho aminoácido, sin que exista ninguna otra restricción. En el eje de ordenadas se representa el número de residuos de la traducción

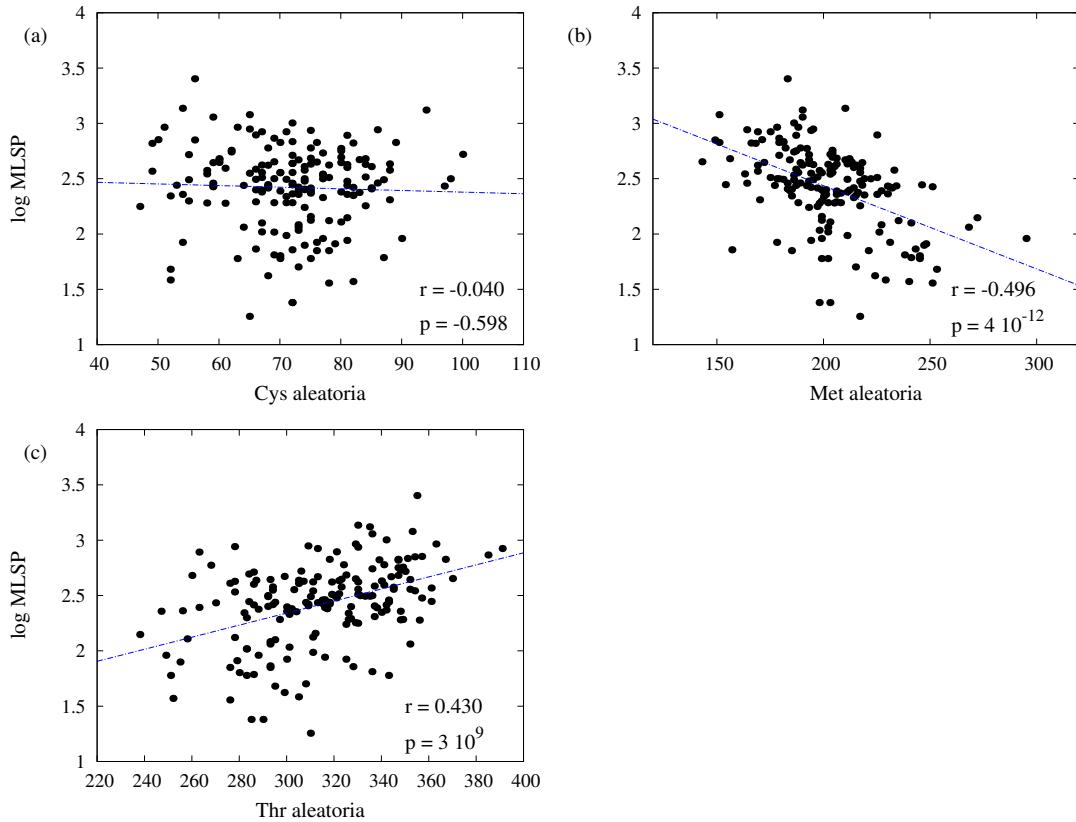


Figura 1.4: Influencia de la composición nucleotídica en la abundancia de cisterna, metionina y treonina en las proteínas codificadas por mtDNA. Para cada especie, se generó una secuencia aleatoria con la misma frecuencia nucleotídica que la secuencia original. Posteriormente, dichas secuencias aleatorias se tradujeron usando el código genético mitocondrial de vertebrados, se representó el número de veces que aparece la cisteína (a), la metionina (b) y la treonina (c) frente a *log MLSP* y se realizó un análisis de correlación para cada par de variables.

de la secuencia original. En el caso de que exista alguna fuerza evolutiva que afecte al contenido del aminoácido, tan solo la distribución en este eje de ordenadas se vería afectada. Dicho de otro modo, el desplazamiento de la nube de puntos con respecto a la bisectriz, pone de manifiesto si existen fuerzas evolutivas distintas del sesgo mutacional que actúan sobre el contenido de estos aminoácidos. En la figura 1.5a se representa esta distribución bidimensional para la Cys. Como se esperaba, se aprecia una gran desviación de la nube de puntos con respecto a la bisectriz, lo que pone de manifiesto una fuerte selección purificadora en contra de este aminoácido.

Por otro lado, la frecuencia de aparición de Thr es similar entre las cadenas real y aleatoria. En la figura 1.5c se observa cómo los resultados se distribuyen con una alta variabilidad a lo largo de la bisectriz, poniendo de manifiesto la ausencia de otra fuerza evolutiva distinta del sesgo mutacional.

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

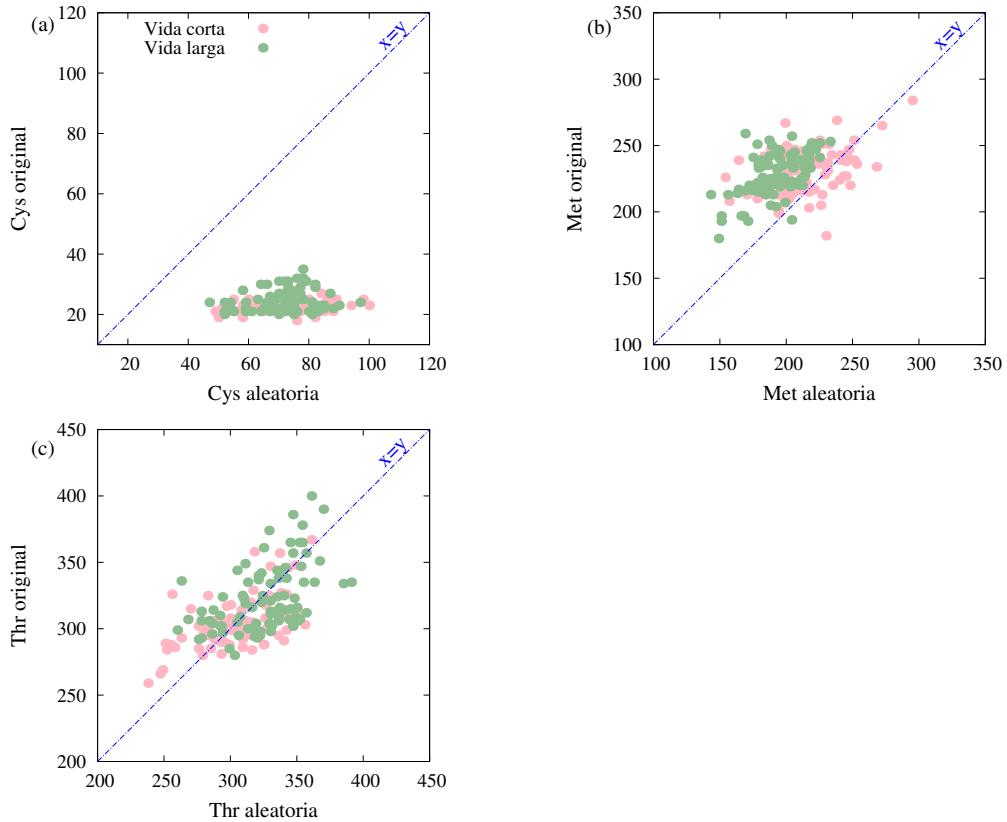


Figura 1.5: Comparación entre la abundancia aminoacídica medida en las proteínas originales y lo esperado por la única influencia de la composición nucleotídica. Para cada una de las 173 especies, se representó el número de cisteínas (a), metioninas (b) y treoninas (c) presentes en las secuencias originales frente al número de estos aminoácidos codificados por una secuencia aleatoria con igual composición nucleotídica. Además, la nube de puntos se dividió en dos grupos: animales de vida corta, cuyo valor de $\log MLSP < 2.46$, y animales de vida larga con $\log MLSP > 2.46$ (véase leyenda). El valor de $\log MLSP = 2.46$ se corresponde con 228.4 meses.

Con respecto a la Met, se observa un comportamiento intermedio entre los descritos para Cys y Thr. Aunque no es tan evidente como en el caso de los residuos Cys, se aprecia un desplazamiento de la distribución con respecto a la bisectriz (figura 1.5b). Este resultado parece indicar que la mayoría de las especies incorporan en sus proteínas codificadas por mtDNA un mayor número de residuos metionil de lo esperado por la única influencia de la composición nucleotídica.

1.5.4 El análisis estadístico de las diferencias entre las secuencias reales y barajadas revela que el contenido de metionina está sometido a selección positiva

Para cada aminoácido, se computó la variable ΔAa como la diferencia entre el número de residuos codificados por la cadena real y la aleatoria (figura 1.6a). En el caso de ΔThr , se observa una distribución centrada en cero, lo que indica que existen pocas restricciones sobre este aminoácido distintas de la frecuencia nucleotídica. Por el contrario, el valor de ΔCys toma valores negativos en todos los casos, poniendo nuevamente de manifiesto la alta selección purificadora a la que está sometida la Cys. Por último, la distribución de ΔMet , está distribuida en torno a 30, con sólo algunas especies cuyos valores son negativos, lo que indica un cierto grado de selección positiva, que favorece la incorporación de metionina en las proteínas llevadas a estudio.

Con el objetivo de aportar resultados respaldados por una alta significación, se llevó a cabo una prueba de signo para las variables ΔThr y ΔMet . Así, se formuló la siguiente hipótesis nula: en ausencia de restricciones evolutivas, la variable tipificada sigue una distribución normal, $Z \sim N(0,1)$. Mientras que en el caso de ΔThr se acepta la hipótesis nula ($z = 0.0$ y $p = 0.5$), para ΔMet se rechaza claramente ($z = 9.2$ y $p = 2 \times 10^{-20}$). No obstante, dado que la mayoría de los genes que codifican para estas proteínas comienzan la traducción con el codón AUG, que codifica para metionina, nuestros resultados podrían estar sesgados por esta causa, por lo que se repitió la prueba habiendo eliminado previamente el codón de inicio en todas las secuencias. De nuevo los resultados para ΔMet fueron estadísticamente significativos ($z = 6.6$ y $p = 2 \times 10^{-11}$) y se pudo rechazar H_0 (figura 1.6b). Estos resultados sugieren que, efectivamente, las proteínas codificadas por el genoma mitocondrial incorporan más residuos de metionina de los esperados por la influencia neutral del sesgo nucleotídico.

1.5.5 Existe una correlación entre la adición de metionina (ΔMet) y la longevidad

Con los datos representados en la figura 1.5b, se ha puesto en evidencia la presencia de una fuerza selectiva que aumenta el contenido de metionina por encima de lo esperado por el único efecto del sesgo nucleotídico. No obstante, este fenómeno no se observa en igual magnitud en animales con distinta longevidad. Desplazando el foco de atención a aquellos puntos, representados en la figura 1.5, que se sitúan por debajo de la bisectriz, se observa que todos ellos, excepto uno, pertenecen a la categoría de animales de vida corta ($MLSP < 2.46$). Esto nos sugiere que debe existir una diferencia entre animales longevos y no longevos en cuanto a la magnitud de la selección positiva a la que está sometida la metionina.

Para investigar esta hipótesis, se llevó a cabo un análisis de correlación entre ΔAa y la longevidad ($\log MLSP$). Tal y como se observa en la figura 1.7, se obtuvo una

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

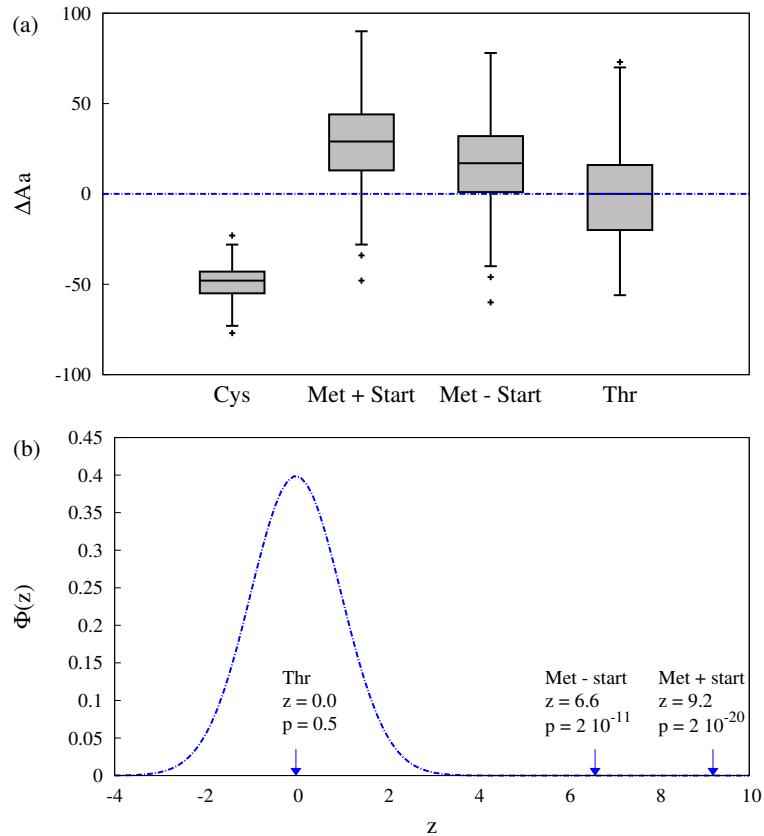


Figura 1.6: Las proteínas codificadas por mtDNA de mamíferos incorporan más residuos metionil que lo esperado por el efecto de la composición nucleotídica. Para cada especie, se computó la diferencia entre el número de veces que aparece cada residuo en la secuencia original y en la aleatoria. La distribución de dichas diferencias (ΔA_a) se representa en (a). Mientras que ΔA_a se distribuye en torno a cero, indicando la ausencia de restricciones selectivas, ΔCys y ΔMet se alejan de lo esperado por azar. Debido a que la mayoría de los genes codificantes usan como codón de inicio el triplete AUG, que codifica para metionina, llevamos a cabo los mismos análisis tanto incluyendo dicho triplete (Met + Start), como eliminándolos de cada gen (Met - Start). En (b) se representan los resultados de la prueba de signo realizada sobre las variables ΔA_a tipificadas, z (véase texto para más detalles), así como una distribución normal dibujada con una línea discontinua. El valor del estadístico Z , calculado para Met + Start y Met - Start fueron de 9.2 y 6.6, respectivamente. Ambos valores resultaron ser significativamente mayores que cero, al contrario que en el caso de Thr, para la que no pudo rechazarse la hipótesis nula.

correlación positiva muy significativa ($r = 0.456$, $p = 3 \times 10^{-10}$) para el caso de ΔMet , mientras que no se observó correlación en los casos de ΔCys y ΔThr ($p = 0.3$ y 0.5 , respectivamente).

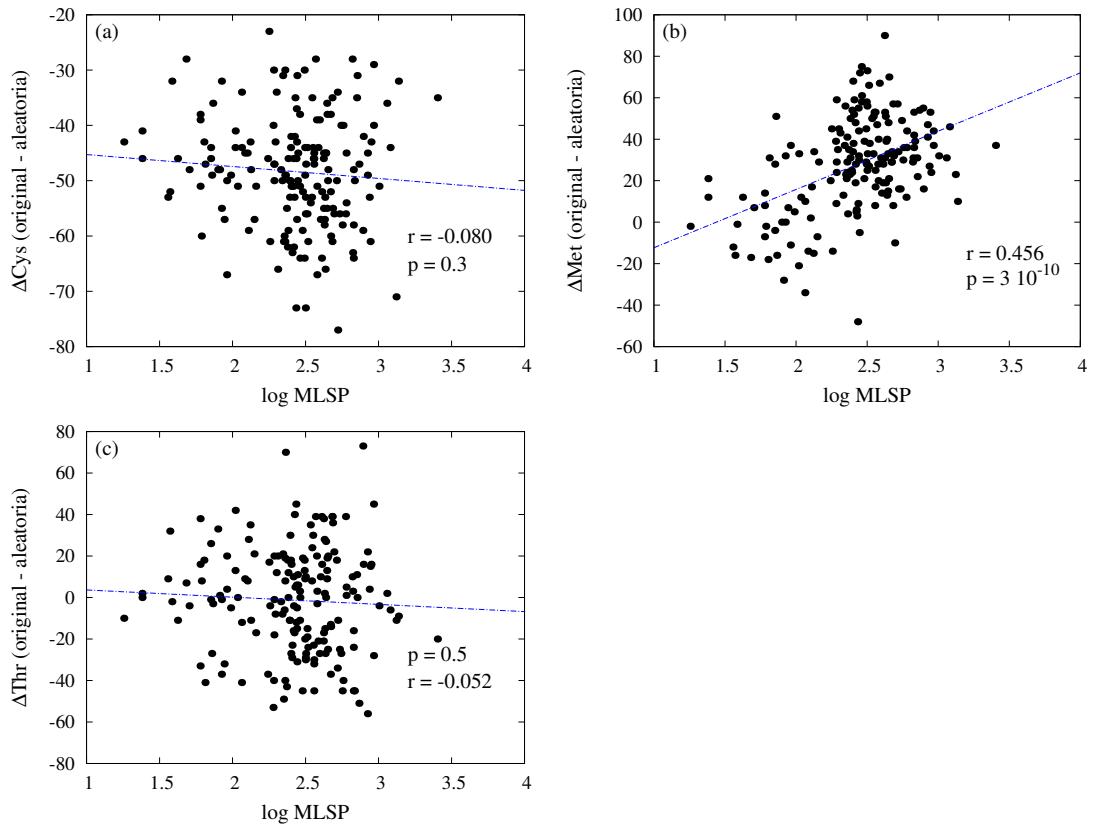


Figura 1.7: Los mamíferos de vida larga añaden metionina activamente en sus proteínas codificadas en el genoma mitocondrial. Se analizaron las correlaciones entre ΔAa y longevidad, cuyos resultados se representan en el interior de cada gráfica.

1.6 Discusión

Recientes estudios realizados en numerosos grupos de animales, ponen de manifiesto correlaciones entre la abundancia de distintos aminoácidos y la longevidad. Sus autores, asocian la pérdida gradual del contenido de cisteína y metionina, y el incremento del uso de treonina con la longevidad, y sugieren un rol adaptativo de estos cambios frente a estrés oxidativo (??; ??) y a favor de un aumento en la estabilidad de la estructura (??; ??). Bender y colaboradores (?), revelaron la existencia de un proceso adaptativo que incrementa la abundancia de metionina en proteínas mitocondriales en defensa ante el estrés oxidativo. En el mismo año, otros autores realizaron estudios similares con respecto a la treonina (??; ??), en los que también se sugería la presencia de mecanismos adaptativos relacionados con la mejora de la estructura de proteínas integrales de membrana. Jobson y colaboradores (?), en completo desacuerdo con estas hipótesis, realizaron distintos análisis con una muestra mayor de varios grupos de especies. Por un lado, encontraron una abundancia de metionina menor en animales más longevos (menor estrés oxidativo).

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

Según estos autores, este es un patrón que contradice las teorías de Bender y colaboradores. Además, sugieren que el cambio en la abundancia de metionina se debe a la pérdida de codones AUA y AUG debido a la reducción de timina (T), sin tener en consideración ningún otro proceso evolutivo distinto del sesgo mutacional (?). Con respecto a la treonina, en la que se observa una fuerte dependencia con la longevidad, Jobson y colaboradores (?) también eluden la adaptación como fuerza evolutiva, adjudicando este rol únicamente a la composición nucleotídica.

En nuestra opinión, la existencia de un mecanismo que explique la dinámica de un fenómeno evolutivo, no implica que no existan otros procesos que también influyan en el proceso. Así, el trabajo de investigación con el que comienza esta tesis, insiste en la posibilidad de que existan fuerzas selectivas que contribuyan a la variabilidad de la proporción de metionina en el genoma mitocondrial, así como la evolución de otros dos residuos de interés, la cisteína y la treonina.

1.6.1 Una nueva aproximación al estudio de la influencia del sesgo mutacional sobre la frecuencia de aminoácidos

Como punto de partida, y con los datos genéticos disponibles para este trabajo, se observó la correlación entre longevidad y frecuencia de nucleótidos en los genes de la hebra ligera del genoma mitocondrial (figura 1.3). Seguidamente, se analizó la relación entre la frecuencia nucleotídica y la cantidad de residuos de cisteína, metionina y treonina codificados por secuencias aleatorias con la misma composición nucleotídica que las originales (figura 1.4). Hasta donde sabemos, el barajado de secuencias (véase Material y Métodos para más detalle) es un método no aplicado antes en esta línea de investigación. Así, en este estudio se aborda de un modo distinto la influencia del sesgo mutacional sobre la variación en el contenido de aminoácidos de las proteínas con origen en la mitocondria. En principio, los resultados vuelven a poner de manifiesto la estrecha relación entre composición nucleotídica y abundancia de aminoácidos, por lo que puede considerarse una nueva evidencia a favor de la hipótesis de Jobson et al. (?). No obstante, empleando el modelo de trabajo descrito anteriormente, se han realizado nuevos análisis que han permitido aislar los efectos del sesgo mutacional de otras fuerzas evolutivas. Por un lado, la comparación del número de residuos codificados por las secuencias reales y barajadas (figura 1.5), pone de manifiesto la tendencia general de cada aminoácido estudiado, en la que se observan desviaciones con respecto al efecto neutral de la composición nucleotídica, que resultaron ser significativas para cisteína y metionina (figura 1.6). Por otro lado, se observó una dependencia entre el incremento de residuos de metionina (ΔMet) y la longevidad de las especies que componen la muestra (figura 1.7). Estos resultados, que se discuten de aquí en adelante, complementan la teoría propuesta por Jobson et al. (?), sugiriendo que, además del sesgo mutacional, existen fuerzas selectivas que promueven el aumento de residuos de metionina en proteínas codificadas por el genoma mitocondrial, haciendo frente a la negación categórica de la presencia de procesos adaptativos en este escenario.

1.6.2 La cisteína es excluida de las proteínas, pero la variación en su abundancia no depende de la longevidad

La selección purificadora en contra de la cisteína no es discutible hoy día. Éste es el aminoácido menos frecuente en proteínas y su abundancia apenas varía entre especies (?). Además, el hecho de que todos los puntos se acumulen por debajo de la bisectriz en la figura 1.5a, sugiere que, efectivamente, existe una presión purificadora en contra del uso de este residuo. Sin embargo, no se ha observado una diferencia significativa en cuanto a la desviación del efecto del sesgo mutacional entre especies con distinta longevidad. Esto significa que, aunque es evidente que los residuos cisteínil están excluidos, las especies longevas y no longevas los evitan de igual forma. Esta afirmación contrasta con las hipótesis expuestas en el trabajo de Moosmann y Behl (?), en el que observaron una correlación negativa entre longevidad y contenido de cisteína en proteínas mitocondriales (?). En esta línea, es necesario apuntar ciertos aspectos metodológicos que pueden explicar las diferencias en las conclusiones de Moosman y Behl y los resultados de este trabajo. Aledo et al. (?), sugieren que la dependencia observada entre longevidad y abundancia de residuos de cisteína es debido a un artefacto, puesto que ambas variables son dependientes de la filogenia. Estos autores reprodujeron la misma correlación en un conjunto de especies de mamíferos, más acotado que los anteriores, pero la dependencia desaparece al corregir para la inercia filogenética (?; ?), lo cual sí está en línea con las observaciones reportadas en esta tesis.

1.6.3 La variación del contenido de treonina está influenciada exclusivamente por el sesgo mutacional

La variación observada en la cantidad de residuos de metionina y treonina codificados por las secuencias que componen nuestra muestra, ilustra el efecto del sesgo mutacional sobre la abundancia de estos residuos. Estos resultados son más acusados en el caso de la treonina, en el que los puntos están distribuidos de manera uniforme alrededor de la bisectriz de la figura 1.5. Esto sugiere que el contenido de este residuo en proteínas mitocondriales, está modulado directamente por la composición nucleotídica de los genes codificantes. Así, los animales más longevos acumulan mayor cantidad de residuos treonil. Esta relación fue expuesta por Kitazoe et al. (?) y Min y Hickey (?) en estudios realizados casi simultáneamente. Según sus autores, este incremento responde a un proceso de adaptación, dado que una mayor frecuencia de treonina favorece la formación de hélices transmembrana, por lo que contribuye a mejorar la estabilidad de los componentes de la cadena respiratoria, dando lugar a mayor esperanza de vida para el organismo. Sin embargo, aunque nuestros resultados sí evidencian que existe una correlación entre el contenido de treonina y longevidad, no se observan efectos de una fuerza selectiva que varíe la frecuencia de su aparición, ya que no se obtuvo un valor de ΔThr significativamente distinto de cero (figura 1.6), es decir, no existe una desviación del efecto neutral del sesgo mutacional. Este argumento está en línea con las conclusiones que alcanzaron

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

Jobson y colaboradores (?) con respecto a la treonina, quienes presentaron teorías contrarias al escenario adaptativo propuesto por Kitazoe et al. (?). Además, estas evidencias sirven como modelo de referencia en el que se acepta la hipótesis nula sobre la presencia de fuerzas evolutivas distintas del sesgo mutacional (figura 1.6b).

1.6.4 La variación del contenido de metionina está impulsada por mecanismos selectivos además del sesgo mutacional

Centrados en el caso de la metionina, sí se encuentran evidencias de selección positiva sobre la abundancia de este residuo. En primer lugar, se observa que toda la nube de puntos está desplazada hacia la parte superior de la bisectriz en la figura 1.5b. Además, la media de la diferencia entre el número de residuos de metionina codificados en la secuencia original y los esperados por la única influencia de la composición nucleotídica (ΔMet) es significativamente mayor que cero (figura 1.6a y b), incluso habiendo eliminado los codones de iniciación (AUG) de las secuencias. Estos resultados sugieren que las proteínas codificadas por el genoma mitocondrial incorporan más residuos metionil que lo esperado por la influencia neutral del sesgo mutacional.

1.6.5 Los animales más longevos incorporan más metioninas en sus proteínas de lo esperado por la composición nucleotídica de sus genes

Es interesante prestar atención a las especies que se sitúan por debajo de la bisectriz en la figura 1.5b. Éstas son aquellas que no incorporan más metioninas de lo esperado por azar y se corresponden, por tanto, con las que se distribuyen por debajo de la línea del efecto de la composición nucleotídica (cero) en la figura 1.6a. La mayoría son especies de baja longevidad, lo cual indica que en éstas, no es necesaria la presencia de mecanismos que aumenten el contenido de metionina distintos del sesgo mutacional.

Ante un escenario hipotético en el que cualquier especie de mamíferos necesite una misma cantidad de residuos metionil en sus proteínas, independientemente de su longevidad, aquellas especies que presenten una composición nucleotídica desfavorable (menor número de tripletes que codifican para este residuo), necesitarían otros mecanismos distintos del sesgo mutacional, que permitieran la incorporación de metionina. Esta hipótesis se ajusta a las evidencias expuestas en este trabajo. Dado que los animales más longevos presentan una mayor frecuencia de timinas en la hebra sentido del genoma mitocondrial (figura 1.3), los codones ACA y ACG (que codifican para treonina) son más frecuentes que AUA y AUG (que codifican para metionina). Así, éstos estarían en desventaja frente a los animales de vida corta ante una necesidad equivalente de residuos de metionina. En la línea de esta hipótesis (los animales de vida larga incorporan más metioninas que los de vida corta con respecto al esperado por la composición nucleotídica), los resultados de la correlación entre longevidad y ΔMet (figura 1.7b) arrojaron

un nuevo argumento a su favor. A mayor longevidad, más activa es la incorporación de residuos metionil a las proteínas codificadas por mtDNA. Este comportamiento ha sido hallado sólo en el caso de metionina.

1.6.6 La ausencia de correlación entre longevidad y ΔCys y ΔThr se explica por distintas causas

Los valores de ΔCys y ΔThr no exhiben un comportamiento similar al de ΔMet , por lo que se puede afirmar que no existe ningún mecanismo selectivo dependiente de la longevidad que modifique la frecuencia de estos aminoácidos. No obstante, es probable que la ausencia de correlación para estos dos residuos sea debida a situaciones evolutivas distintas. Por un lado, es evidente que el contenido de treonina es dependiente de la composición nucleotídica de forma exclusiva, por lo que la ausencia de fuerzas selectivas hace que los valores de ΔThr estén en torno a cero, independientemente de la longevidad de la especie.

Por otro lado, con respecto al comportamiento de ΔCys , existe una remarcable abolición de residuos de cisteína en todas las proteínas en general, lo cual ha sido atribuido al rol pro-oxidante de este aminoácido, especialmente cuando se encuentra en la superficie (?). Bien es sabido que el daño oxidativo provoca un declive funcional asociado con el envejecimiento en animales (?), por lo que sería esperable que ΔCys presentara dependencia con la longevidad, siendo las especies sometidas a un mayor estrés oxidativo aquellas que presentaran valores más negativos (mayor presión purificadora), sin embargo, la pendiente en la correlación entre longevidad y ΔCys no es significativamente distinta de cero. Esta aparente incongruencia puede ser razonada por una situación en la que se ha alcanzado el mínimo número posible de residuos cisteinil en el proteoma. La funcionalidad de muchos residuos de cisteína, tales como aquellos que están implicados en la unión con hierro en numerosos sitios de la cadena respiratoria, hacen indispensable la incorporación de este aminoácido en determinadas posiciones. Por estos motivos, no sería posible una anulación completa del contenido de este residuo, sino que se encuentra en un límite inferior infranqueable, donde ya no se aprecia dependencia con la longevidad.

1.6.7 No se observa selección purificadora en contra de la metionina en animales de vida larga, sino una presión adaptativa en animales de vida corta

Tal y como hemos señalado repetidas veces en este capítulo, la correlación entre longevidad y abundancia de metionina es negativa, es decir, que los animales más longevos presentan menores frecuencias de este residuo en sus proteínas (?; ?; ?). Sin embargo, la interpretación evolutiva de este hecho permanece aún discutida. Por una parte, unos autores afirman que existe una selección negativa en contra de este residuo en animales

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

de vida larga (?). Éstos argumentan que juega un papel pro-oxidante y por ello los animales longevos tienden a disminuir la sensibilidad de su proteoma frente a ROS, lo que les confiere una ralentización del envejecimiento. Otros, ya mencionados anteriormente, aportan una visión neutralista de esta correlación, en la que proponen que la variación en el contenido de metionina se debe al sesgo mutacional, al igual que ocurre con otros aminoácidos como la treonina (?). Por último, recientes estudios y entre ellos, este trabajo de tesis, aportan argumentos a favor de una selección positiva de residuos metionil en animales de vida corta (?). Estos últimos se apoyan en la teoría de que la metionina puede actuar como sumidero de ROS y evitar la oxidación de residuos circundantes como el triptófano (?; ?), y posteriormente reducirse en una reacción catalizada por la metionina sulfóxido reductasa (?; ?). Así, cada metionina oxidada y posteriormente reducida, elimina un equivalente de ROS, evitando que degrade de forma irreversible a otros componentes.

El modelo de trabajo que se ha seguido en esta tesis, ofrece resultados que contrastan con la teoría de la selección purificadora propuesta por Ruiz et al. (?), puesto que no se observa una desviación significativa de los valores esperados por la composición nucleotídica en animales de vida larga. De ser así, los puntos de la figura 1.5b se distribuirían por debajo de la bisectriz, como en el caso de la cisteína (figura 1.5a) y la correlación entre longevidad y ΔMet sería negativa. En este punto, se propone que existe una influencia del sesgo mutacional sobre el contenido de metionina, que sumado a una presión adaptativa en animales de vida larga (que presentan una composición nucleotídica desfavorable), constituyen las fuerzas evolutivas que determinan la cantidad de residuos metionil en proteínas codificadas por el genoma mitocondrial. Esta evidencia de la presencia de selección positiva, está en línea con los argumentos a favor del papel anti-oxidante de la metionina.

1.6.8 La abundancia de metionina en proteínas puede estar determinada por un compromiso entre su capacidad de eliminar ROS y el efecto nocivo de su oxidación

La controversia actual frente a la dualidad del rol pro- o anti-oxidante de la metionina, puede ser abordada desde un punto de vista imparcial, siendo quizás la que sigue, la teoría más parsimoniosa alcanzada hasta el momento en la literatura. Tal y como se ha observado para el caso de la cisteína, en el que su frecuencia se encuentra en un estado estacionario, acotado por su funcionalidad en un lado y por su sensibilidad a daño oxidativo en otro, puede ocurrir de un modo similar, aunque más distendido, con la metionina. Un estudio reciente ha demostrado que el enzima metionina sulfóxido reductasa (MSR), el principal elemento en el que se basan los argumentos a favor del papel anti-oxidante de la metionina (?; ?; ?), es más eficiente sobre proteínas desplegadas y nacientes (?). A partir de esta observación, se deduce que un aumento en el contenido de residuos metionil en las proteínas, supondrá una mayor defensa ante daño oxidativo mediado por ROS. Además, una elevada expresión de los genes codificantes, tal y como

ocurre en la mitocondria (?), puede aumentar la concentración de sustrato disponible para MSR, y así constituir un mecanismo eficiente en la eliminación de ROS de la matriz. Por otra parte, no deben obviarse los efectos nocivos de la oxidación de este residuo a metionina sulfóxido (MetO), unidos al hecho de que no todos los residuos metionil pueden ser sustratos de MSR (?), independientemente de si la proteína está plegada o no. Por ejemplo, la oxidación de dos residuos de metionina en la calmodulina impide la unión proteína-proteína debido a la incompatibilidad de MetO en alfa-hélices estables (?). Otro caso similar se da en la hormona de crecimiento humana, que incrementa su susceptibilidad *in vitro* a desnaturalización por temperatura por la oxidación de dos metioninas de su superficie (?). Ante este escenario se puede considerar que, para una proteína concreta, un enriquecimiento en metionina supone, por una parte, un mecanismo de eliminación de ROS beneficioso para toda la mitocondria, y por otra, un elemento potencialmente desestabilizador para la propia proteína, que causaría por consiguiente, una pérdida total o parcial de su función.

1.7 Conclusiones

1. Existen procesos adaptativos que aumentan el contenido de residuos de metionina en proteínas codificadas por mtDNA.
2. La selección positiva detectada sobre el contenido de metionina tiene más relevancia en animales de vida larga.
3. Existe una correlación positiva entre la magnitud de la selección positiva sobre la abundancia de metionina y la longevidad.
4. El cambio en la abundancia de metionina a lo largo de la evolución se ajusta a un papel anti-oxidante de este residuo.
5. La presión purificadora sobre residuos de Cys tiene igual fuerza independientemente de la longevidad. Esto es debido a que la abundancia de este residuo está reducida al mínimo posible para mantener la integridad de la estructura y/o función de las proteínas analizadas.
6. No existen procesos adaptativos que modifiquen el contenido de treonina en proteínas codificadas por el genoma mitocondrial, sino que está influenciado exclusivamente por el sesgo mutacional. De este modo, el comportamiento evolutivo de la treonina ha servido, además, como modelo de referencia en estos análisis.

1. Fuerzas adaptativas sobre el contenido de metionina en mtDNA

Capítulo 2

Citocromo b y COX 1 presentan una dinámica evolutiva diferencial condicionada por la estabilidad termodinámica

2.1 Resumen

Combinando información evolutiva y estructural de 231 especies de mamíferos, hemos abordado la influencia de determinantes estructurales sobre la tasa evolutiva de citocromo b y COX 1, dos proteínas codificadas por el genoma mitocondrial. Los residuos del interior, no expuestos al solvente, de citocromo b, en contraste con los de COX 1, exhiben una notable tolerancia a los cambios. Esta mayor capacidad evolutiva¹ de citocromo b contrasta con la menor tasa de sustituciones sinónimas de su gen, en comparación con el de COX 1, sugiriendo que este último está sujeto a una fuerte selección purificadora. En este trabajo presentaremos resultados que apuntan a que el efecto sobre la estabilidad de las mutaciones ($\Delta\Delta G$), podría ser la causa del comportamiento evolutivo diferencial de estas dos proteínas mitocondriales.

2.2 Introducción

Las mitocondrias, además de desempeñar un papel central en la fosforilación oxidativa (OXPHOS), están involucradas en diversos procesos celulares tales como crecimiento y

¹Entiéndase *capacidad evolutiva* o *evolvabilidad* como la predisposición de un gen a experimentar cambios en su secuencia a lo largo de la evolución. Es una traducción adaptada del término en inglés «*evolvability*»

2. Dinámica evolutiva de citocromo b y COX 1

proliferación (?), apoptosis (?) y envejecimiento (?). No resulta sorprendente, pues, que alteraciones en la biología mitocondrial hayan sido relacionadas con numerosas enfermedades (?). Estas alteraciones pueden venir provocadas por mutaciones en el genoma mitocondrial (mtDNA) o en genes nucleares (nDNA) que codifican para proteínas mitocondriales (?; ?). Como hemos visto en la introducción general, el genoma mitocondrial de mamíferos codifica tan sólo para 13 proteínas de OXPHOS (figura 2), mientras que el grueso de proteínas mitocondriales están codificadas por genes nucleares. Sin duda, la evolución del mtDNA presenta numerosas particularidades cuando la comparamos con la evolución del nDNA. A ello contribuyen las características propias del genoma mitocondrial que destacamos con anterioridad.

El hecho de que, en vertebrados, el genoma mitocondrial cambia más rápidamente que el genoma nuclear, parece estar, actualmente, fuera de discusión. Sin embargo, este hallazgo supuso una gran sorpresa (?), dado que parecía desafiar la idea bien asentada de que cuanto más importante es la función de una proteína, menor es el ritmo al que cambia su estructura primaria (?). Desde entonces, la identificación de los factores que determinan el ritmo al que una proteína evoluciona, ha atraído una considerable atención.

Trabajos recientes sugieren que todos los eventos que influyen sobre la expresión de una proteína, tales como la transcripción, *splicing* y traducción, poseen una profunda influencia sobre la velocidad a la éstas evolucionan (?; ?). Así, pues, se acepta que el nivel de expresión de un gen es uno de los principales determinantes de la velocidad a la que evoluciona su correspondiente proteína (?). Sin embargo, cuando analizamos el ritmo al que evolucionan los distintos residuos de una misma proteína, hemos de desplazar el foco de atención hacia aspectos estructurales y/o funcionales. En este sentido, las propiedades funcionales de una proteína, incluyendo interacciones con otras proteínas y modificaciones postraduccionales, están todas relacionadas con las propiedades de la superficie de dicha proteína. Sin embargo, la habilidad de la molécula para plegarse correctamente y la estabilidad termodinámica de dicha conformación, la cual en última instancia determina la función, están fuertemente influenciadas por las características del interior de la proteína (?; ?). No es de extrañar, pues, que el grado de exposición al solvente de un residuo, sea una de las propiedades estructurales que ha recibido particular atención como posible determinante de la evolución proteica.

Estudios realizados en levaduras (?; ?; ?) y bacterias (?), apoyan la visión de que los residuos enterrados en el núcleo de la proteína, son más refractarios a los cambios evolutivos cuando se comparan con su contrapartida de residuos expuestos en la superficie. Ante estas observaciones, uno podría sentirse tentado a especular que aquellas proteínas con una menor proporción de residuos expuestos deberían evolucionar lentamente. Aunque tal razonamiento ha recibido apoyo de parte de algunos autores (?), otros han documentado observaciones opuestas. Esto es, que proteínas con una baja proporción de residuos expuestos evolucionan más rápidamente (?). Para explicar esta observación, aparentemente poco intuitiva, Franzosa y Xia (?) han sugerido que un aumento del núcleo proteico tiene poca influencia sobre la evolución de los residuos ente-

2.3. Objetivos

rrados, pero el aumento en la estabilidad termodinámica podría relajar las restricciones impuestas sobre los residuos expuestos, que podrían tolerar, entonces, un mayor rango de sustituciones.

Dado que, como hemos brevemente documentado en esta introducción, el efecto de la accesibilidad de un residuo al solvente sobre la capacidad evolutiva de dicho residuo ha sido materia de debate, consideramos interesante abordar esta cuestión empleando proteínas codificadas por el genoma mitocondrial, ya que éstas a menudo muestran patrones evolutivos distintos a los encontrados en proteínas codificadas por nDNA (?). Para tal propósito, hemos explorado la dinámica evolutiva tanto de los residuos de superficie como de los residuos enterrados de citocromo b y COX 1, dos proteínas codificadas por mtDNA, comúnmente empleadas en estudios filogenéticos.

2.3 Objetivos

A la luz de los antecedentes anteriores, decidimos plantear los siguientes objetivos concretos:

1. Describir y comparar la variabilidad de cada posición de la estructura primaria, en relación con su accesibilidad al solvente.
2. Estimar la contribución de cada residuo a la estabilidad termodinámica de la conformación nativa de la cadena polipeptídica, así como a la estabilidad del complejo enzimático del cual forma parte dicha cadena.
3. Determinar si existe una relación entre la variabilidad de cada residuo y su contribución a la estabilidad termodinámica.

2.4 Material y métodos

2.4.1 Datos y modelado molecular

Se obtuvieron 231 genomas mitocondriales de mamíferos (figura 2.1) de la base de datos *Genome*, perteneciente al NCBI (*National Center for Biotechnology Information*, <http://www.ncbi.nlm.nih.gov>). Este conjunto de especies llevado a estudio comprende 27 órdenes distintos (véase apéndice A para más detalle). Los alineamientos de múltiples secuencias se llevaron a cabo con versiones superiores a las 2.0 de ClustalW (?), configurado con los siguientes parámetros: 15 y 6.6 para penalizaciones por *gap* y extensión, respectivamente; 30 % para el retraso de secuencias divergentes² y 0.5 para el

²Traducido de «*delay divergent sequences*»

2. Dinámica evolutiva de citocromo b y COX 1

ponderado de transiciones³. La identidad de las secuencias para citocromo b y COX 1 fueron de 73 y 90 %, respectivamente.

Se generaron por homología modelos de las estructuras de citocromo b y COX 1 para las secuencias ortólogas, usando como plantillas las estructuras obtenidas experimentalmente por difracción de rayos X correspondientes a *Bos taurus*. Éstas se obtuvieron de la base de datos PDB (*Protein Data Bank*) con los identificadores 1be3 y 2occ para citocromo b y COX 1, respectivamente (véase apéndice F, tabla F.8). Los cálculos estructurales se realizaron en Swiss-Model (?; ?; ?). Debido a dificultades en la generación de modelos fiables para todas las especies, los siguientes análisis estructurales se realizaron sobre 221 estructuras de citocromo b y 189 de COX 1. En la figura 2.1 se detalla para qué especies se obtuvieron dichos modelos.

2.4.2 Determinación de las posiciones enterradas y expuestas

La superficie expuesta al solvente de una proteína fue definida por primera vez en 1971 por Lee y Richards (?). En 1973, Shrake y Rupley desarrollaron el algoritmo de la bola rodante para realizar el cálculo, que consiste en una simulación en la que una sonda esférica con un radio determinado (generalmente 1.4 Å, que coincide con el radio de una molécula de agua) recorre toda la superficie del modelo. En este trabajo se usó Surface Racer 5.0 (?), que devolvía la accesibilidad de cada residuo, definida ésta como el porcentaje de superficie expuesta del residuo en cuestión (%ASA). Para computar tal valor, se toma como referencia (100 %), la superficie expuesta del aminoácido examinado en un tripéptido del tipo G-X-G en su conformación extendida ($\Phi = -120^\circ$, $\Psi = 140^\circ$), donde G es glicina y X es el aminoácido estudiado (?). En la tabla A.2 (apéndice A) se detallan los valores de superficie máxima en angstroms de cada aminoácido.

Cada residuo i se clasificó en base al valor de su %ASA $_i$, considerándose enterrado si es menor al 5 %, y expuesto en caso contrario ($\geq 5\%$). Así, tras un alineamiento múltiple, se computó el número de veces que aparece cada residuo enterrado o expuesto en el conjunto de modelos atómicos obtenido anteriormente. Una posición se consideró enterrada cuando expone menos del 5 % de la superficie del aminoácido en la mayoría de las especies ($>50\%$).

2.4.3 Determinación de la entropía de Shannon

El análisis de la entropía de Shannon es una herramienta diseñada para cuantificar la diversidad de un sistema. Para los alineamientos múltiples de citocromo b y COX 1, se

³Traducido de «*transition weight*»

2.4. Material y métodos



Figura 2.1: Árbol filogenético de las especies de mamíferos usadas en este estudio. Se obtuvo una colección de 231 especies almacenadas en las bases de datos del NCBI. Para cada especie, se concatenaron las secuencias de citocromo b y COX 1, se alinearon y se calculó la topología del árbol filogenético mediante el programa PHYLP. Se ha anexado un vector bidimensional al nombre de cada especie. La primera coordenada indica si existe (1) o no (0) un modelo estructural confiable para citocromo b, y la segunda, se refiere de igual forma a COX 1.

2. Dinámica evolutiva de citocromo b y COX 1

computó la variabilidad de cada posición mediante la ecuación 2.1.

$$H_c(j) = - \sum_{i=1}^c p_i(j) \log_c p_i(j) \quad (2.1)$$

donde $p_i(j)$ es la frecuencia del residuo perteneciente a la clase i en posición j , y c es el número de clases. Se calcularon dos valores de entropía para cada posición j : $H_6(j)$, cuando los 20 aminoácidos se agrupan en 6 categorías ($c = 6$) según sus propiedades físico-químicas y sus patrones naturales de sustitución (?), y $H_{20}(j)$, en el que $c = 20$ y, por tanto, los aminoácidos no se agrupan. La entropía toma valores comprendidos entre 0 y 1; un valor de entropía $H_c(j)$ bajo significa que existe poca variabilidad en la posición j , si toma un valor nulo, significa que la posición es invariante o, dicho de otro modo, en todas las secuencias del alineamiento encontramos en la posición j un aminoácido perteneciente a la misma clase. En caso opuesto, los valores cercanos a 1 implican una alta variabilidad en esa posición. Estos cálculos se llevaron a cabo mediante un script en Perl programado *ad hoc*.

2.4.4 Clasificación de posiciones en base a sus valores de entropía de Shannon

Según los valores de entropía, y atendiendo a los siguientes criterios, se distinguieron cuatro tipos de posiciones: (i) valores bajos de H_6 y H_{20} caracteriza la posición como constreñida, puesto que se han permitido pocas sustituciones en la evolución. Contrariamente, (ii) las posiciones con valores altos de ambas entropías se han nominado como no constreñidas. (iii) Valores bajos de H_6 y altos de H_{20} implica que las sustituciones encontradas en esa posición ocurren entre aminoácidos con propiedades físico-químicas similares, y se categorizan como posiciones conservativas. Por último, (iv) valores bajos de H_{20} pero altos de H_6 , indica que hay pocas sustituciones pero, la mayoría de ellas, se dan entre aminoácidos pertenecientes a clases distintas. Esta última categoría puede reflejar cambios adaptativos dirigidos por selección positiva.

Para determinar cuándo un valor de $H_c(j)$ es alto o bajo, se calcularon, tras excluir previamente los residuos invariantes ($H_{20}(j) = 0$), los cuartiles superior (UQ_c) e inferior (LQ_c), que se emplearon como umbrales superior e inferior, respectivamente (figura 2.2).

2.4.5 Cálculo de tasas de sustituciones

El cálculo de las tasas de sustituciones empleando todos los pares de especies que se pueden formar a partir de la colección de 231 especies que disponemos (figura 2.1), puede no ser fiable debido a la variación entre especies de algunas características de la secuencia como la frecuencia nucleotídica. De hecho, como ya se ha explicado en

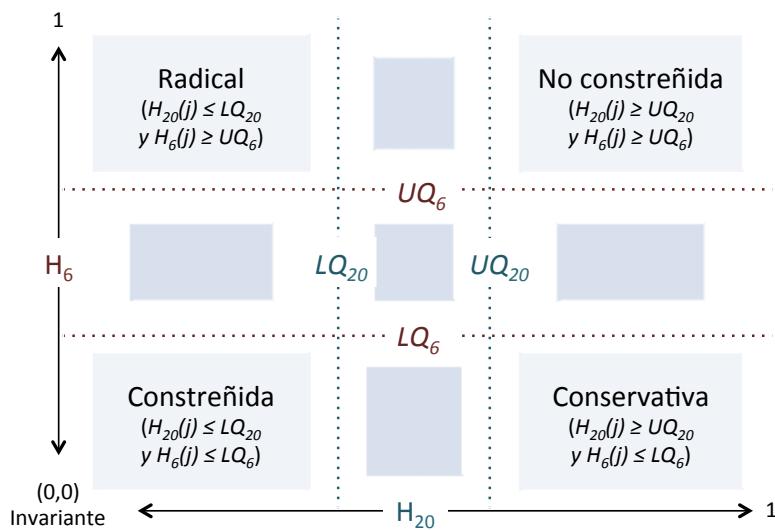


Figura 2.2: Esquema de clasificación de posiciones según los valores de entropía de Shannon. Para cada proteína (citocromo b y COX 1), se calcularon los valores de los cuartiles superior UQ_c e inferior LQ_c para cada valor de entropía H_6 y H_{20} . Los valores de los cuatro cuartiles de citocromo b fueron: $UQ_6 = 0.0650$, $LQ_6 = 0.0049$, $UQ_{20} = 0.0551$ y $LQ_{20} = 0.0079$. Para COX 1, los valores de los cuartiles fueron $UQ_6 = 0.0222$, $LQ_6 = 0.0040$, $UQ_{20} = 0.0376$ y $LQ_{20} = 0.008$.

varias ocasiones en esta memoria, el sesgo nucleotídico puede emular el efecto de la selección positiva (?). Por este motivo, tan solo calculamos las tasas de sustituciones entre pares de secuencias pertenecientes a especies filogenéticamente próximas (véase apéndice D). Para identificar todos los pares de especies cercanas presentes en nuestra muestra, se construyó un árbol filogenético a partir de secuencias de dos genes concatenados que codifican para RNA ribosómico, en lugar de usar las mismas secuencias de citocromo b y COX 1, de esta forma, se evitaba la posibilidad de un comportamiento tautológico. La resolución de la filogenia se realizó mediante métodos de máxima verosimilitud implementados en el programa PHYLIP (*the PHYLogeny Inference Package*, <http://evolution.genetics.washington.edu/phylip/general.html>). Para el cálculo de las tasas de sustituciones, sólo se consideraron adecuados los pares formados por especies que se encuentran unidas directamente al mismo nodo interno. Se obtuvieron, entonces, 54 pares de especies cercanas. A partir de las secuencias de citocromo b y COX 1 se calculó en número medio de sustituciones nucleotídicas por sitio mediante el método de Nei-Gojobori (?) y se aplicó la corrección de Jukes-Cantor para inferir múltiples sustituciones en el mismo sitio. Así, para cada gen (citocromo b y COX 1), se obtuvieron 54 puntos en el plano $d_S \times d_N$, donde d_S representa el número medio de sustituciones sinónimas por sitio sinónimo y d_N es la media de las sustituciones no sinónimas por sitio no sinónimo.

A continuación, cada alineamiento de cada uno de los genes analizados, se dividió en dos nuevos alineamientos segregando cada triplete según la accesibilidad al solvente del

2. Dinámica evolutiva de citocromo b y COX 1

residuo para el que codifica. Así, se obtuvieron dos alineamientos por cada gen, uno con los tripletes que codifican para posiciones enterradas y otro con triplets que codifican para posiciones expuestas. Para cada alineamiento, se calculó de nuevo d_S y d_N siguiendo la misma metodología explicada anteriormente.

2.4.6 Análisis de la estabilidad termodinámica

Los cálculos de estabilidad termodinámica de los modelos atómicos estudiados en esta tesis, se realizaron mediante métodos predictivos de alta fiabilidad, implementados en el algoritmo FOLDEF e integrado en el programa FOLD-X (?). Tal y como explican sus autores, la estabilidad de una proteína se expresa como la diferencia de la energía libre (ΔG) entre el péptido plegado y desnaturalizado.

Las estructuras 3D de citocromo b y COX 1 fueron sometidas en primer lugar a un proceso de optimización integrado en FOLD-X llamado “reparación” (comando *RepairPDB*), que identifica aquellos residuos que tienen ángulos de torsión incorrectos o colisiones de *van der Waal* (véase manual de usuario de FOLD-X para más detalle sobre el funcionamiento de este comando). Posteriormente, se ejecutó el comando *alascan*, que sustituye cada residuo por alanina y devuelve el valor de $\Delta\Delta G$ ($\Delta G_{silvestre} - \Delta G_{mutante}$) de cada simulación. Este proceso dio como resultado dos matrices de 221 x 379 y 189 x 514 con los valores de $\Delta\Delta G$ para citocromo b y COX 1, respectivamente.

2.4.7 Análisis estadísticos

Las distribuciones aleatorias se generaron mediante sencillos *scripts* en Perl. Los cálculos de probabilidad se realizaron con el programa Wolfram Mathematica 8.0. El resto de los análisis estadísticos se hicieron con SPSS 15.0.

2.5 Resultados

Para abordar la cuestión de si la superficie e interior de las proteínas mitocondriales están sujetas a las mismas constricciones evolutivas, comenzamos por designar como “enterrada” o “expuesta” a cada posición de la estructura primaria de citocromo b y COX 1. Para tal propósito se empleó el criterio de accesibilidad expuesto en Materiales y Métodos. Brevemente, para cada una de las proteínas ortólogas de citocromo b, y otro tanto para las de COX 1, se obtuvieron modelos 3D. Esto permitió computar el área de superficie expuesta (ASA) de cada residuo en cada proteína ortóloga. La accesibilidad al solvente de un residuo dado, fue calculada como la razón entre la ASA del residuo en la estructura nativa de la proteína y la ASA de dicho residuo en una cadena polipeptídica desnaturizada y extendida ($\Phi = -120^\circ$, $\Psi = 140^\circ$). De esta forma, tras realizar un alineamiento múltiple, cada posición fue designada como expuesta si en la mayoría de las

especies ($>50\%$) el residuo que ocupa dicha posición tenía una accesibilidad superior al 5 %, en caso contrario la posición se consideró enterrada (figura 2.3).

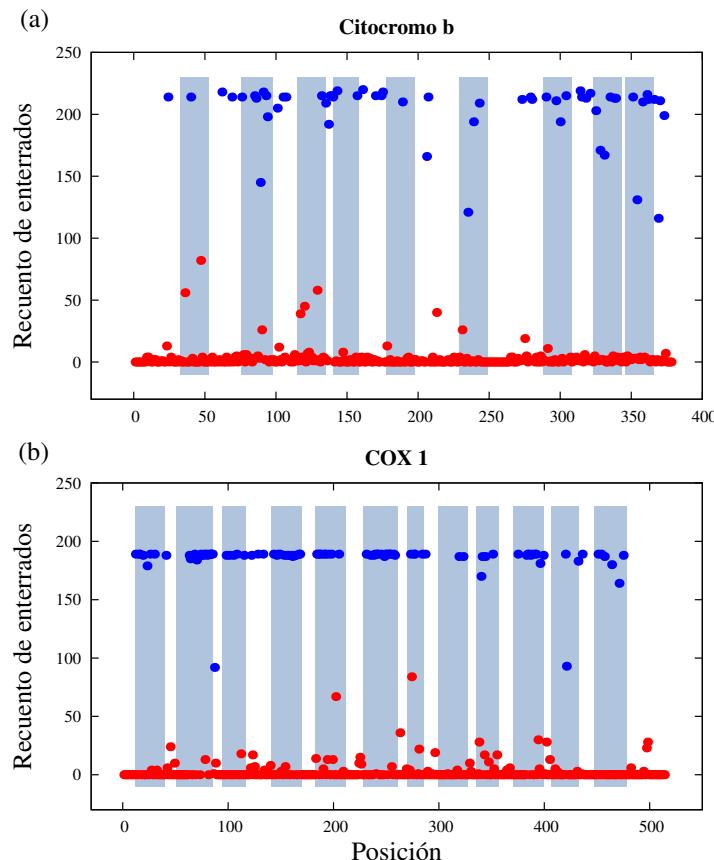


Figura 2.3: Discriminación entre residuos expuestos y enterrados. Se determinó la accesibilidad de cada residuo en cada uno de los modelos atómicos de citocromo b y COX 1, obtenidos para cada especie. Los aminoácidos que exponen más del 5 % de su superficie se clasificaron como expuestos y el resto como enterrados. Se computó el numero de instancias (especies) en las que cada residuo aparece enterrado y se representó gráficamente frente a la posición del residuo en la estructura primaria de citocromo b (a) y COX 1 (b). Los rectángulos verticales de color abarcan las posiciones designadas en *UniProt* como transmembranales. Aquellas posiciones que aparecen enterradas en la mayoría de las especies ($>50\%$) han sido consideradas como posiciones interiores (círculos azules). Las posiciones restantes se han considerado como posiciones de superficie (círculos rojos).

2. Dinámica evolutiva de citocromo b y COX 1

2.5.1 El interior de COX 1 está enriquecido en posiciones invariantes, pero éstas se distribuyen aleatoriamente a lo largo de toda la estructura primaria de citocromo b

En una primera aproximación, una vez que cada posición se encontraba adscrita bien al conjunto de residuos enterrados o bien al de residuos expuestos, decidimos centrar nuestra atención en aquellos residuos que han permanecido invariantes durante la diversificación de los mamíferos. Más concretamente, formulamos la siguiente hipótesis nula: las posiciones invariantes se distribuyen aleatoriamente entre el interior y la superficie de las proteínas consideradas. Para contrastar esta hipótesis, definimos una variable aleatoria, X , como el número de residuos invariantes que se encuentran enterrados en el interior de la proteína. Para cada una de las dos proteínas analizadas, computamos el valor que toma X , la proporción de residuos invariantes (p_x = número de residuos invariantes/número total de residuos) y el número de residuos enterrados (n). En la tabla 2.1 se resumen tales datos. De esta forma, bajo las condiciones definidas por la hipótesis nula, la variable aleatoria X sigue una distribución binomial, $X \sim \text{Bin}(n, p_x)$. Esto nos permitió calcular la probabilidad de que, por azar, encontráramos un número de residuos invariantes enterrados igual o superior al observado. Aunque tal probabilidad fue baja en el caso de ambas proteínas, citocromo b y COX 1, sólo en la última pudimos rechazar la hipótesis nula con un nivel de significatividad del 1% ($p\text{-valor} = 0.009$).

Tabla 2.1: Abundancia de residuos expuestos, enterrados e invariantes en citocromo b y COX 1.

Proteína	P_{exp}	p_x	n	x	$P[X \geq x]$
Citocromo b	0.8482	0.314	58	22	0.175
COX 1	0.6718	0.556	170	110	0.009

P_{exp} es la proporción de residuos expuestos, p_x es la proporción de residuos invariantes, n es el número de residuos enterrados, x es el número de residuos invariantes enterrados y $P[X \geq x]$ es la probabilidad de encontrar por azar un número de residuos invariantes enterrados mayor o igual al observado.

2.5.2 El interior de COX 1, pero no el de citocromo b, exhibe una entropía de Shannon por debajo de lo justificable por el azar

Hasta ahora, tan sólo hemos analizado la distribución de posiciones invariantes, considerando una posición como invariante cuando el mismo aminoácido se encuentra, sin excepción, en las 231 especies analizadas. Para explotar la información contenida en el resto de posiciones, decidimos adoptar el formalismo de la teoría de la información para estudiar el grado de conservación evolutiva del interior de las dos proteínas mitocon-

2.5. Resultados

driales objeto de este estudio. De esta forma, la variabilidad de cada posición puede ser medida mediante la llamada entropía de Shannon (véase la sección metodológica, ecuación 2.1).

Tras computar H_6 y H_{20} para cada posición, se calculó la media para los residuos enterrados. Así, los valores medios de $H_6(\text{enterrado})$ y $H_{20}(\text{enterrado})$ fueron 0.023 y 0.019, respectivamente, para citocromo b. Estos valores disminuyeron considerablemente al considerar COX 1, $H_6(\text{enterrado}) = 0.006$ y $H_{20}(\text{enterrado}) = 0.005$. Puesto que la entropía de Shannon es una variable que nos cuantifica el grado de conservación de una posición, estábamos en disposición de falsar la siguiente hipótesis nula: los residuos enterrados no están más conservados que el resto de residuos. Para tal propósito, comparamos las entropías medias observadas para los residuos enterrados $H_c(\text{enterrado})$, con la distribución de los valores medios de entropía de un conjunto de residuos tomados aleatoriamente de la proteína, y del mismo tamaño que el conjunto de residuos enterrados, $H_c(\text{aleatorio})$. Así pues, para citocromo b se formaron 10^5 conjuntos aleatorios de 58 residuos cada uno, mientras que para COX 1 obtuvimos 10^5 conjuntos aleatorios de 170 residuos cada uno. Para cuantificar la significatividad de las diferencias observadas entre la entropía del interior de una proteína y la entropía de conjunto aleatorio de residuos, se cuantificó el número de veces que ocurrió $H_c(\text{aleatorio}) < H_c(\text{enterrado})$ después de llevar a cabo los 10^5 experimentos. Este dato, convenientemente expresado, nos proporciona el error tipo I (o error tipo alfa), que nos permite estimar la probabilidad de observar por azar un valor medio de entropía menor que el observado entre los residuos enterrados. De esta forma, con un nivel de significatividad $\alpha=10\%$ no pudimos rechazar la hipótesis nula para el caso de citocromo b. Por el contrario, en el caso de COX 1 la hipótesis nula fue rechazada con errores tipo alfa tan bajos como $\alpha = 0.34\%$ y 0.00% , para H_6 y H_{20} , respectivamente. Por tanto, estos resultados están en línea con los descritos anteriormente para los residuos invariantes.

Dado que observamos que las posiciones enterradas de COX 1 están enriquecidas con residuos invariantes, podríamos pensar que los bajos valores de $H_c(\text{enterrado})$ que hemos calculado anteriormente, son debidos fundamentalmente a la contribución de los residuos invariantes del núcleo de COX 1. En otras palabras, deseábamos averiguar si, una vez excluidos del análisis las posiciones invariantes, seguíamos concluyendo que los residuos enterrados están más constreñidos que los residuos expuestos en la superficie. Para abordar tal cuestión, una vez excluidas las posiciones invariantes, generamos conjuntos aleatorios de 36 y 60 residuos de citocromo b y COX 1, respectivamente. Estos conjuntos se utilizaron para generar distribuciones de $H_c(\text{aleatorio})$. La figura 2.4 muestra el resultado de semejante análisis. Como se puede deducir de dicha figura, los residuos variables enterrados de COX 1 se encuentran mucho más constreñidos que el resto de posiciones variables de COX 1 ($\alpha < 5\%$). Sin embargo, para citocromo b, las diferencias entre posiciones enterradas y expuestas no alcanzaron significatividad estadística ($\alpha > 18\%$).

2. Dinámica evolutiva de citocromo b y COX 1

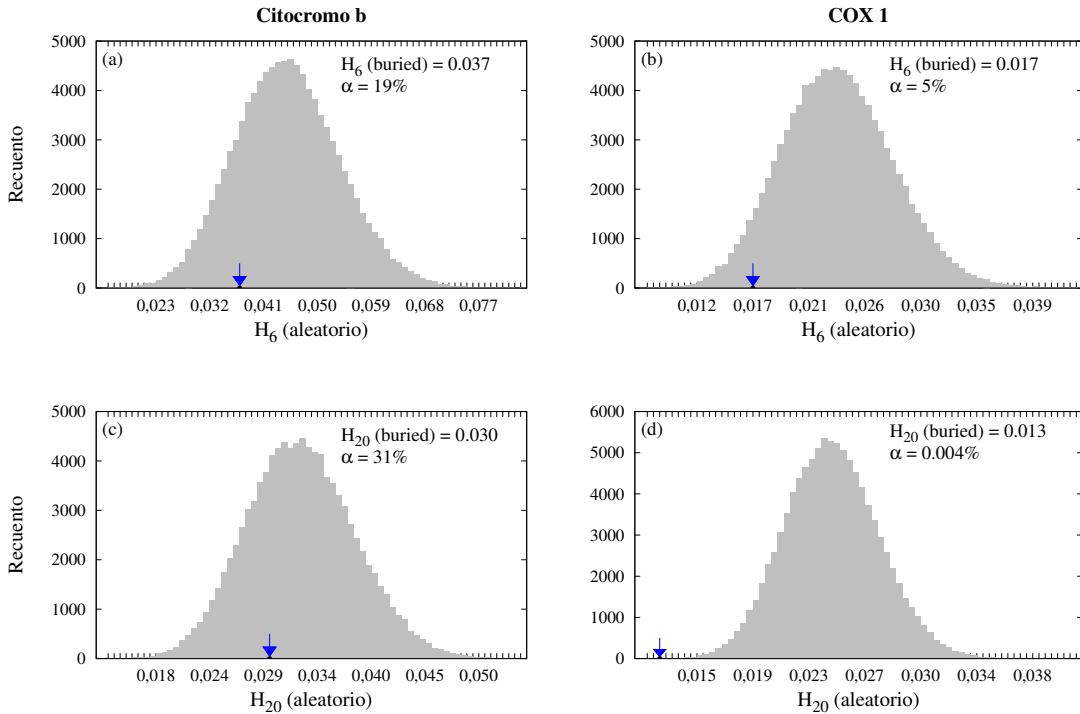


Figura 2.4: Entropía de Shannon de las posiciones enterradas de citocromo b y COX 1. Una vez excluidas las posiciones invariantes, se calcularon los valores medios de entropía (H_6 y H_{20}) de las posiciones enterradas, representados en la esquina superior derecha de cada gráfica y con una flecha en el eje de abscisas. Estos valores se compararon con una distribución de valores medios de entropía de un número igual de residuos escogidos al azar de cualquier región de la proteína (interior y superficie).

2.5.3 Las posiciones no constreñidas se distribuyen aleatoriamente a lo largo de citocromo b, pero son selectivamente excluidas del interior de COX 1

Para cada proteína se formaron tres conjuntos de posiciones designados *No constreñido*, *Conservativo* y *Radical*, siguiendo el criterio expuesto en Materiales y Métodos (figura 2.5). Una vez definidos y formados estos conjuntos de posiciones, decidimos analizar si los residuos pertenecientes a cada una de estas categorías se encuentran presente preferentemente en, o por el contrario tienden a ser excluidos de, el interior de las proteínas. Para este propósito, computamos las frecuencias de residuos no constreñidos ($p_u = 0.068$ y 0.054), conservativos ($p_c = 0.018$ y 0.037) y radicales ($p_r = 0.034$ y 0.031) en citocromo b y COX 1, respectivamente. Estas frecuencias se emplearon como estimadores de las probabilidades de que dada una posición cualquiera, ésta pertenezca a la correspondiente categoría. Por otro lado, definimos tres variables aleatorias (U , C , R) como el número de posiciones no constreñidos enterradas, conservativas enterradas y radicales enterradas,

respectivamente. Si estos tres tipos de posiciones se distribuyen aleatoriamente entre el interior y la superficie de las proteínas, cabría esperar que cada una de estas variables aleatorias siguiera una distribución binomial: $U \sim Bi(n, p_u)$, $C \sim Bi(n, p_c)$, $R \sim Bi(n, p_r)$, siendo n el número de residuos enterrados en la proteína bajo consideración. De esta forma, después de computar los valores que estas variables aleatorias toman en citocromo b y COX 1, estábamos en condiciones de conocer la probabilidad de encontrar por azar un número igual o superior a éste observado en la proteína bajo análisis. Una vez calculadas estas probabilidades (tabla 2.2), concluimos que el interior de estas proteínas mitocondriales no está particularmente enriquecido en ninguno de estos tres tipos de posiciones, ya que en todos los casos la probabilidad anteriormente referida superaba el valor de 0.01, que es el límite de significatividad que habíamos fijado.

A continuación, examinamos la posibilidad alternativa de que alguna de estas tres categorías de posiciones estuviera siendo selectivamente excluida del interior de las proteínas. Para tal propósito, calculamos, asumiendo una distribución de residuos aleatoria, la probabilidad de encontrar en el interior de la proteína, debido únicamente al azar, un número de residuos de la categoría bajo examen igual o inferior al observado. De esta forma, cuando las probabilidades calculadas quedan por debajo de 0.01, rechazamos la hipótesis nula (los residuos de la categoría considerada se distribuyen aleatoriamente), y concluimos que el interior de la proteína excluye a este tipo de residuo. Como podemos observar en la tabla 2.2, tan sólo el interior de COX 1 parece excluir selectivamente los residuos de la categoría no constreñida.

Tabla 2.2: Probabilidades, de acuerdo con la hipótesis nula, de encontrar en el interior de cada proteína un número de posiciones no constreñidas (u), conservativas (c) o radicales (r) menor o mayor que el observado.

Proteína	n	$P[U \leq u]$	$P[C \leq c]$	$P[R \leq r]$	$P[U \geq u]$	$P[C \geq c]$	$P[R \geq r]$
Cit. b	58	0.438	0.913	0.135	0.764	0.281	1.000
COX 1	170	0.008**	0.05	0.101	0.991	0.983	0.969

n es el número de residuos enterrados de cada proteína

Las hipótesis nulas asumen que el número de posiciones no constreñidas, conservativas y radicales siguen una distribución binomial (véase texto para más detalle)

** Nivel de confianza menor que 0.01

2.5.4 El gen de COX 1 está sujeto tanto a una presión selectiva como a una selección purificadora mayores que las del gen de citocromo b

Bajo la asunción de que las mutaciones silenciosas (sinónimas) son fundamentalmente neutras (?), la comparación de las velocidades de sustituciones sinónimas, d_S , entre los genes de COX 1 y citocromo b, nos sugiere una sorprendente mayor presión mutacional para COX 1 (figura 2.6a). Por el contrario, cuando la variable comparada es la velocidad

2. Dinámica evolutiva de citocromo b y COX 1

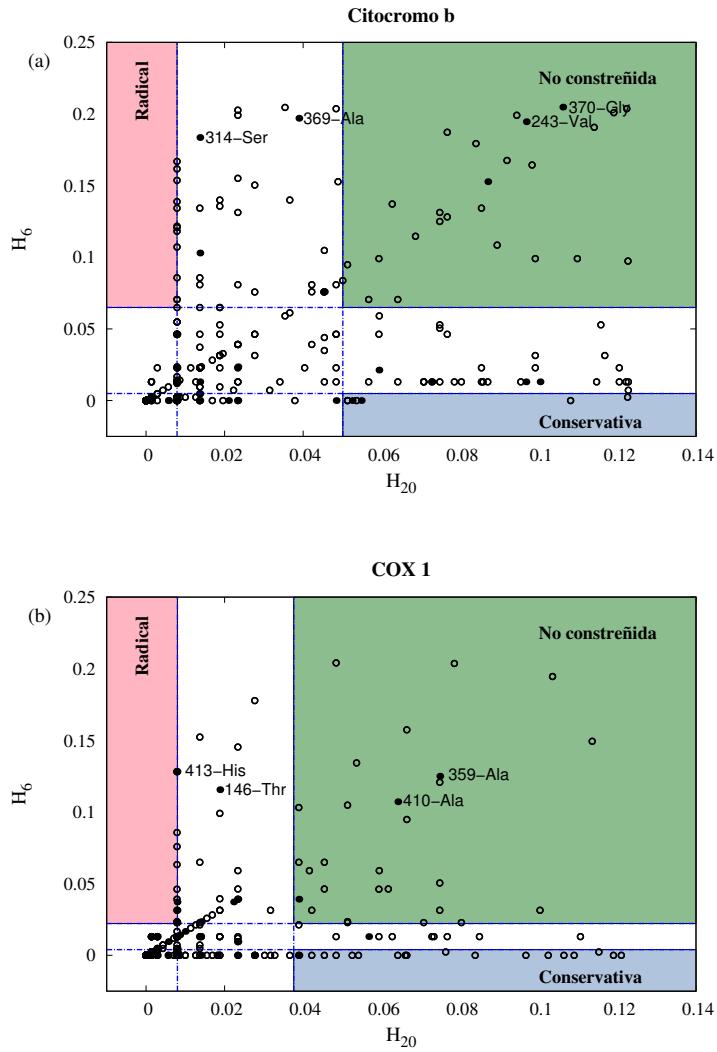


Figura 2.5: Identificación y localización de las posiciones no constreñidas, conservativas y radicales. Para cada proteína, se representa H_6 frente H_{20} . Las líneas verticales y horizontales se corresponden con los los cuartiles superior e inferior de cada variable. Los círculos llenos se corresponden con aquellas posiciones que se encuentran enterrados, mientras que los círculos vacíos representan las posiciones expuestas.

de sustituciones no sinónimas, d_N , COX 1 mostró un menor valor, indicando que este gen está sujeto a un mayor grado de restricción debido a una fuerte selección purificadora (figura 2.6b).

En este punto nos preguntamos si habría diferencias en la presión mutacional y en la selección purificadora, entre los residuos expuestos y enterrados de una misma proteína. Como se muestra en las figuras 2.6c y d, no observamos diferencias significativas en los

2.5. Resultados

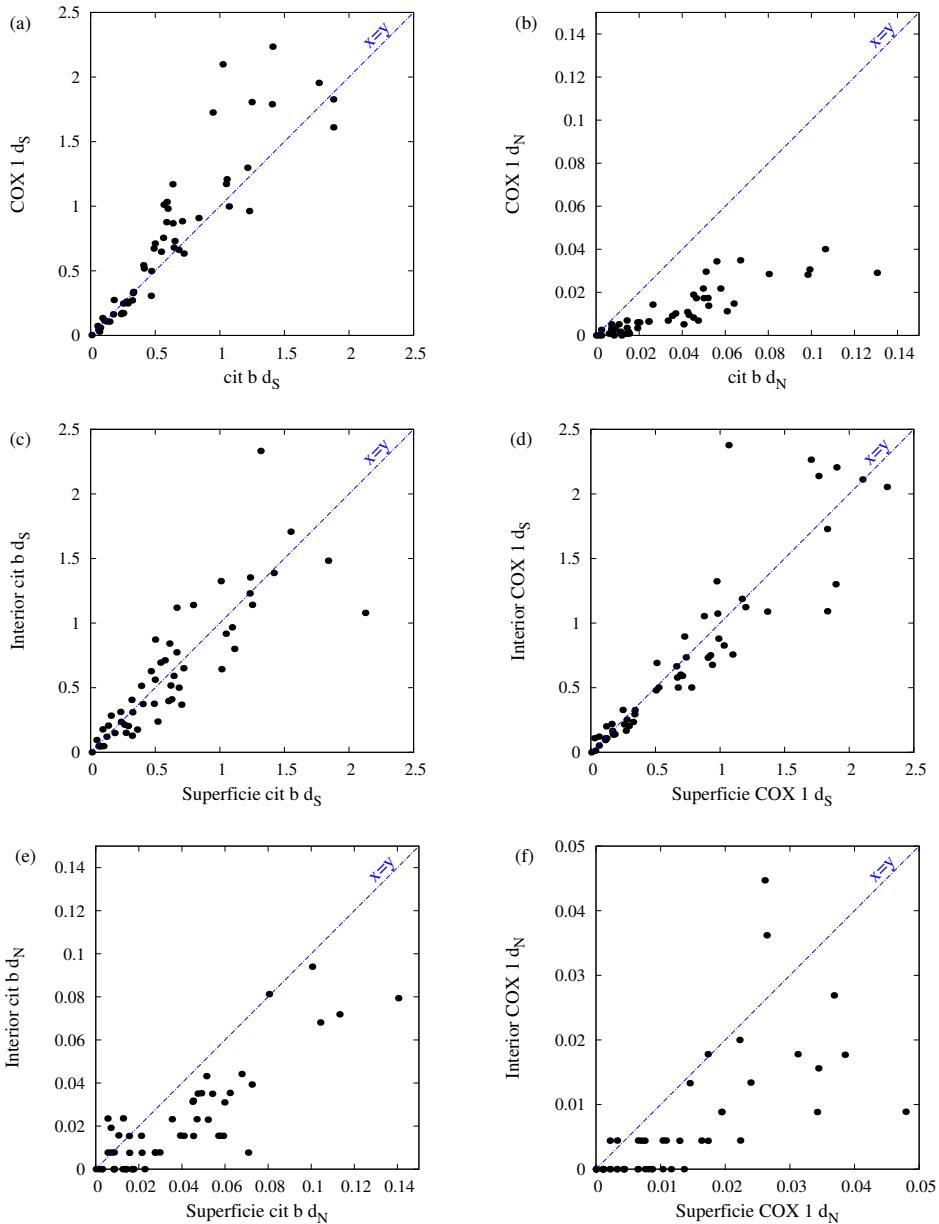


Figura 2.6: Comparación de las tasas de sustitución inter- e intragénicas. Se calcularon las tasas de sustitución por sitio para cada par de especies próximas usando distintos conjuntos de residuos de citocromo b y COX 1 (véase texto para más detalles), y se representaron gráficamente para comparar los resultados.

valores de d_S del interior y la superficie de las proteínas. Por el contrario, las velocidades de sustituciones no sinónimas fueron notablemente menores entre los nucleótidos que codificaban para aminoácidos en posiciones enterradas, independientemente del gen

2. Dinámica evolutiva de citocromo b y COX 1

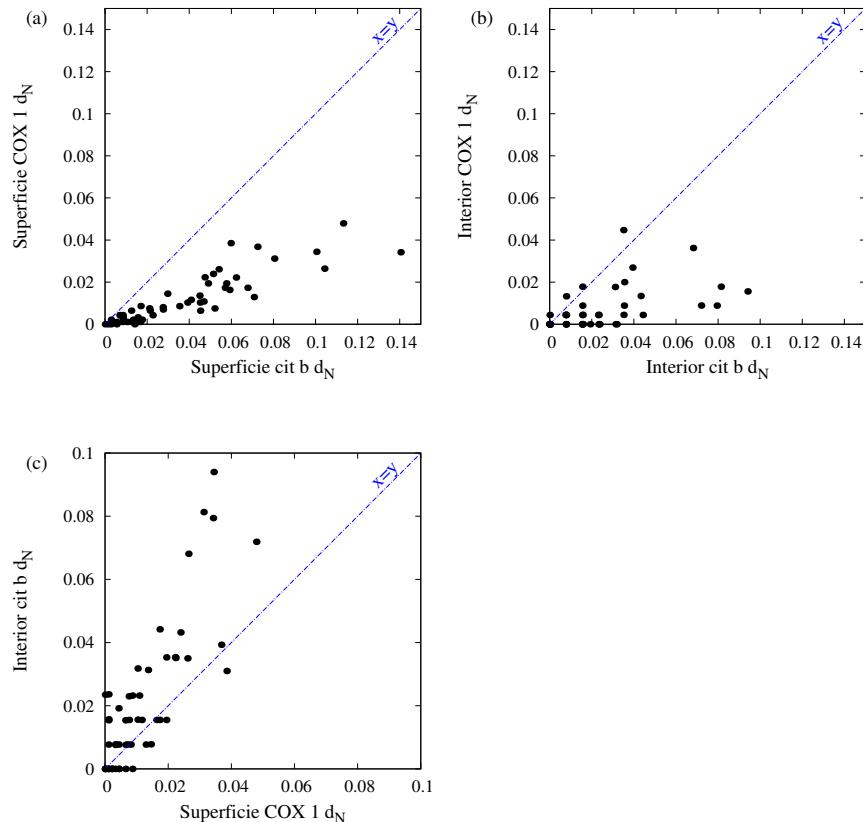


Figura 2.7: Comparación de tasas de sustituciones no sinónimas entre distintos conjuntos de residuos. Se compararon las tasas de sustituciones no sinónimas por sitio para los conjuntos de residuos expuestos de citocromo b y COX 1 (a). De igual forma, se compararon los resultados entre los residuos del interior de ambas proteínas (b). Por último, se comparó del interior de citocromo b con la superficie de COX 1 (c).

examinado (figuras 2.6e y f). No obstante, hay que destacar que el valor de d_N para posiciones enterradas de COX 1 fue mucho menor que el observado para citocromo b (figura 2.7). Aún así, la velocidad de sustituciones no sinónimas que afecta a los residuos enterrados de citocromo b estuvo por encima del valor d_N correspondiente a los residuos expuestos de COX 1 (figura 2.7c).

2.5.5 La estabilidad termodinámica puede dar cuenta del comportamiento diferencial de COX 1 y citocromo b

A menudo, la sustitución de un simple aminoácido en una proteína puede alterar dramáticamente la estabilidad de la misma. No sorprende, pues, que la estabilidad termodinámica haya sido señalada como un determinante de la capacidad evolutiva de las proteínas (?; ?). De acuerdo con esta visión, postulamos la hipótesis de que las mutaciones que afectan a residuos enterrados de COX 1 tienen un efecto desestabilizante mayor que las que tienen lugar en el interior de citocromo b. Para contrastar esta hipótesis de trabajo, hemos analizado el efecto de estabilidad termodinámica ($\Delta\Delta G$) que conlleva la sustitución individual de cada uno de los residuos enterrados por alanina. Para tal propósito, los cambios de estabilidad termodinámica de estas mutaciones fueron computados usando FoldX (?; ?).

En primer lugar, formulamos la hipótesis nula de que las mutaciones puntuales que afectan a posiciones enterradas no son más desestabilizadoras que las mutaciones puntuales que afecten a un sitio elegido aleatoriamente. Para contrastar esta hipótesis, se calculó la media del cambio de energía libre de la mutación de los residuos enterrados a alanina ($\Delta\Delta G$). Este valor se comparó con la distribución de medias correspondientes a conjuntos de residuos aleatorios del mismo tamaño que el conjunto de residuos enterrados. Para obtener dicha distribución, se tomaron 10^6 conjuntos aleatorios de 58 (para citocromo b) ó 179 (para COX 1) residuos. Cada uno de estos residuos se cambió por alanina y se calculó su correspondiente $\Delta\Delta G$, lo que permitió obtener el valor medio para cada conjunto.

Como cabría esperar, las mutaciones que afectaban a residuos enterrados de COX 1 resultaron ser fuertemente desestabilizantes. La hipótesis nula pudo rechazarse con un valor de α inferior a una millonésima. Las mutaciones que afectaban al interior de citocromo b también fueron significativamente más desestabilizantes ($\alpha = 0.0002$) que las mutaciones al azar. Sin embargo, cuando comparamos las distribuciones de $\Delta\Delta G$ de ambas proteínas, se hace evidente que citocromo b es mucho más robusto frente a mutaciones, desde el punto de vista de la estabilidad termodinámica (figura 2.8). Estas observaciones sugieren que las sustituciones no sinónimas en posiciones enterradas podrían ser más fácilmente toleradas en el caso de citocromo b, mientras que estarían mucho más constreñidas en COX 1, lo cual está en línea con nuestros resultados anteriores (figura 2.7).

2. Dinámica evolutiva de citocromo b y COX 1

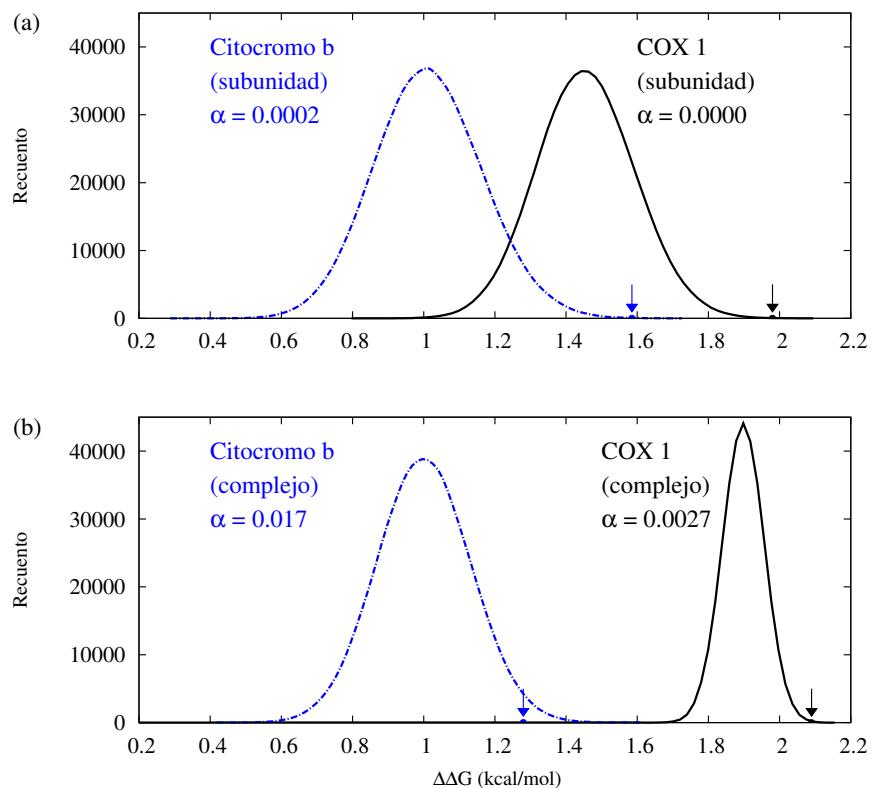


Figura 2.8: Efecto desestabilizador de mutaciones aleatorias en citocromo b y COX 1. El incremento medio en la energía libre ($\Delta\Delta G$) causado por mutaciones en residuos del interior se representa en las gráficas mediante flechas azules (citocromo b) y negras (COX 1). Cada valor se comparó con una distribución de valores medios de $\Delta\Delta G$ obtenidos a partir de conjuntos del mismo tamaño pero compuesto por residuos tomados al azar de toda la proteína (véase texto para más detalles). Dichas comparaciones se realizaron tanto para el caso de la proteína aislada (a) como para la proteína formando parte de su respectivo complejo (b).

2.6 Discusión

En la última década, se han llevado a cabo estudios que ponen de manifiesto la importancia de la superficie expuesta al solvente de una proteína sobre su capacidad evolutiva. Hasta la fecha, se han llevado a cabo experimentos con proteínas solubles en bacterias (?) y en levaduras (?; ?; ?; ?), cuyo resultado común es una mayor variabilidad en aquellos residuos expuestos al solvente sobre los del interior. La concordancia de estos trabajos, sugiere de una forma muy preliminar que esta regla puede hacerse extensible a todas las proteínas. Sin embargo, dada la gran diversidad de estructuras proteicas existentes en la naturaleza, así como los diferentes ambientes con los que interaccionan, es necesario realizar más experimentos que aporten robustez a esta teoría. A priori, no se puede aceptar que proteínas de membrana presenten un comportamiento evolutivo igual a las proteínas solubles con respecto a las diferencias entre superficie e interior. De hecho, ya se ha sugerido anteriormente que las proteínas integradas en la membrana evolucionan de forma distinta de lo que lo hacen otras (?; ?). Por ello, se han escogido para realizar estos análisis, citocromo b y COX 1, que se encuentran en esta categoría. Otra característica muy importante que diferencia nuestra muestra de las analizadas en los estudios precedentes, es el origen genómico. Éstas están codificadas por el genoma mitocondrial de mamíferos, que, como hemos apuntado anteriormente, es un genoma con una estructura y capacidad evolutiva muy particular en comparación con el nuclear y otros genomas mitocondriales como el de aves o plantas (?). Otros dos aspectos señalados son la implicación de estas proteínas en complejos heteroméricos, lo que da lugar a multitud de contactos intermoleculares, y la función en el transporte electrónico y de protones. Los resultados de este trabajo indican que, efectivamente, la superficie de las proteínas estudiadas evoluciona más deprisa que su interior (figura 2.6), por lo que esta regla puede hacerse extensiva al genoma mitocondrial y a un tipo distinto de proteína sometida a las restricciones explicadas anteriormente. Tomando este hecho como punto de partida, el resto de esta investigación estuvo enfocado a discernir cuáles son los mecanismos evolutivos responsables de la divergencia en la capacidad evolutiva de la superficie y el interior de estas proteínas.

2.6.1 Existen diferencias en la dinámica evolutiva de citocromo b y COX 1

Como explicamos, citocromo b, en comparación con COX 1, es más tolerante a los cambios en el interior. Esta premisa ha sido argumentada en este trabajo con una serie de evidencias. Primero, no se observaron diferencias en la distribución de residuos invariantes entre la superficie y el interior de citocromo b (tabla 2.1). Segundo, la entropía de Shannon de los residuos enterrados de citocromo b no es significativamente menor que la media de un conjunto aleatorio de residuos obtenido de la misma proteína (figura 2.4). Tercero, mientras los residuos no constreñidos se excluyen selectivamente del interior de COX 1, no existen diferencias significativas entre el interior y la superficie de citocromo

2. Dinámica evolutiva de citocromo b y COX 1

b (tabla 2.2). Cuarto, aunque los residuos del interior de citocromo b están más constreñidos que los de su superficie, es necesario enfatizar que la tasa de sustituciones no sinónimas de éstos es significativamente mayor que la de la superficie de COX 1 (figura 2.7c). Por último, el efecto sobre la estabilidad de la proteína de las mutaciones en el interior de citocromo b es comparable, e incluso menor, que el efecto de las mutaciones sobre la superficie de COX 1 (figura 2.8). Todos estos resultados nos dejan un escenario con dos elementos sobre los que discutir: por un lado, la descripción desde un punto de vista comparativo de las dinámicas evolutivas de citocromo b y COX 1, y por otro, la estabilidad termodinámica de ambas proteínas, que parece ser, *a priori*, un factor que ejerce una presión selectiva sobre ellas. Con estas piezas, se pretende dilucidar la presencia de procesos evolutivos, adaptativos o no, que expliquen las diferencias entre citocromo b y COX 1.

2.6.2 La presión mutacional es mayor sobre COX 1 que sobre citocromo b

La presión mutacional es un factor que influye en gran medida en la variabilidad de las proteínas. Durante la replicación, la cadena pesada, que codifica para 12 genes de la cadena respiratoria, se encuentra en el estado de hebra sencilla durante un tiempo considerable, viéndose expuesta a un entorno hostil que genera mutaciones por cambios químicos como desaminaciones. Este periodo no es de igual duración para todos los genes del genoma mitocondrial, y se ha observado una correlación positiva entre el lapso en hebra sencilla del gen durante la replicación (D_{ssH}) y el porcentaje de sitios variables del mismo (?). El gen de COX 1 muestra valores de D_{ssH} menores que citocromo b, sin embargo, los resultados de las tasas de sustituciones sinónimas fueron mayores (figura 2.6a). Así pues, el hecho de que COX 1 presente una mayor tasa de sustituciones sinónimas que citocromo b debe explicarse por otros mecanismos distintos del tiempo de replicación (D_{ssH}), como la frecuencia nucleotídica, o el tiempo en hebra sencilla durante la transcripción (?). De hecho, recientes estudios indican que COX 1 se expresa más que citocromo b en mamíferos (?). Todo esto sugiere que la diferencia en la presión mutacional entre estas proteínas puede estar más influenciada por el estrés que sufre la secuencia durante la transcripción que durante la replicación.

2.6.3 Las diferencias entre la evolución de citocromo b y COX 1 adquiere mayor relevancia en procesos post-transcripcionales

Los residuos que se encuentran en el interior y los que se encuentran en la superficie deben estar sujetos a presiones selectivas distintas, tal y como sugieren las tasas de sustituciones de cada región, dado que mientras que las tasas de sustituciones sinónimas son las mismas (figuras 2.6c y d), las de sustituciones no sinónimas difieren considerablemente (figuras 2.6e y f). ¿Cuáles son, entonces, las fuerzas biológicas que se encuentran detrás de este comportamiento diferencial? Atendiendo al ritmo evolutivo de las proteínas completas,

2.6. Discusión

sin distinguir de momento entre regiones distintas, COX 1 está significativamente más conservada que citocromo b. Este hecho ha sido descrito previamente (?) y está en línea con nuestros resultados (figura 2.6b). Recientes trabajos sugieren que las fuerzas selectivas más importantes que explican las diferencias en la dinámica evolutiva de las proteínas en general, se encuentran en procesos post-transcripcionales. Esto se deduce del hecho de que el nivel de transcripción de un gen funciona como un buen predictor de su conservación, dada la correlación negativa existente entre su tasa evolutiva ($\omega = d_N/dS$) y la concentración de mRNA procedente del gen (?; ?). Tanto es así, que el nivel de expresión explica más de la mitad de la variación de las tasas evolutivas (?). COX 1 se expresa más que citocromo b y ambas se expresan más que el resto de las proteínas codificadas en el mtDNA (?). Aunque este hecho *per se* no pueda explicar las diferencias entre distintas regiones de la molécula, sí es posible que, sea cual sea el proceso post-transcripcional determinante de la dinámica evolutiva, ejerza una presión selectiva sobre residuos aislados. Existen tres teorías, no excluyentes entre sí (?), que explican la covarianza entre el nivel de expresión y la tasa evolutiva. La primera es la selección en contra de la toxicidad de proteínas mal plegadas⁴ (?) sobre la que existen diversos estudios empíricos que han apoyado esta teoría (?; ?; ?; ?; ?). La segunda, propuesta por Gout et al. (?) y Cherry (?) argumenta que esta correlación viene dada por un compromiso entre los efectos beneficiosos de una proteína y el coste energético de su expresión. Por último, Yang et al. (2012) demostraron recientemente que la interacción no específica⁵ entre proteínas puede ser otra de las causas, de forma que exista una mayor presión purificadora en contra de residuos que favorezcan este tipo de uniones (?).

Habiendo estudiado estas tres teorías, cabe preguntarse entonces qué aminoácidos o qué regiones se ven más afectadas por los mecanismos selectivos puestos en relevancia en cada una, y comprobar si se ajustan a las diferencias en la dinámica evolutiva de las distintas regiones de citocromo b y COX 1. Según la hipótesis del núcleo de plegado (SFN)⁶, discutida en profundidad por Mirny et al. (?), existe un conjunto de residuos que son responsables, en mayor medida, del correcto plegado de la proteína. Es lógico pensar entonces, que la presión ejercida por el mal plegado, propuesto por la teoría de Drummond et al. (?) sea más acusado en este grupo de residuos. Dado que los aminoácidos que componen el SFN se encuentran generalmente enterrados (?), ésta es una teoría que encaja con nuestros resultados, explicando la mayor conservación del interior de citocromo b y COX 1 con respecto a sus respectivas superficies. No obstante, se deben tener en cuenta las recientes aportaciones que versan sobre los beneficios en la mitocondria de las proteínas desplegadas (?), discutidas previamente en esta tesis (véase sección 1.6.8), que pueden influir de algún modo en la capacidad evolutiva de estas proteínas.

El modelo de la segunda teoría, asume que cada aminoácido contribuye individual-

⁴Traducido de «protein missfolding theory»

⁵Traducido de «protein missinteraction theory»

⁶Traducido de «specific folding nucleus»

2. Dinámica evolutiva de citocromo b y COX 1

mente a la eficacia biológica de la proteína y, por tanto, el efecto de mutaciones deletéreas es proporcional al nivel de expresión de la proteína (?; ?). Este modelo predice que el alto coste energético de un elevado nivel de expresión, debe verse compensado por un fuerte beneficio energético por parte de cada aminoácido que compone la proteína. Es decir, que estarían vetadas todas las sustituciones aminoacídicas que no trajeran consigo una optimización de la función y, consecuentemente, una mayor rentabilidad energética. Apelando a esta teoría, que no ha sido aún demostrada experimentalmente, sería posible explicar la baja variabilidad de citocromo b y COX 1 con respecto al resto de las proteínas de la cadena respiratoria, puesto que son dos proteínas muy implicadas en la función bioenergética más importante en organismos aerobios. No obstante, Drummond et al. (?), demostraron que las proteínas evolucionan a un ritmo no relacionado con la importancia de su función.

Por último, la reciente teoría de Yang et al. (?), la teoría de las interacciones no específicas, predice una alta selección purificadora en los residuos de superficie, puesto que son éstos los que pueden establecer uniones no específicas y generar un efecto tóxico en la célula, al mismo tiempo que una reducción en la concentración de las proteínas implicadas. Según esta hipótesis, los residuos hidrofóbicos son los más excluidos de la superficie, debido a que son los más proclives a formar uniones inespecíficas. Sin embargo, la proporción de estos residuos en la superficie de citocromo b y COX 1 es mayor que en el interior (datos no mostrados), lo cual es esperable debido a que la función de ambas proteínas se lleva a cabo formando parte de complejos heteroméricos. Este hecho, indica que debe existir un compromiso entre el correcto ensamblaje de citocromo b y COX 1 en los complejos III y IV, respectivamente, y la evasión de las uniones inespecíficas.

Sean o no ciertas estas teorías, las tres pueden explicar de un modo ponderado el comportamiento evolutivo de las proteínas. Así, es muy probable que sean aplicables en distinta medida según sea el objeto de estudio. En el caso de citocromo b y COX 1, las teorías del plegado incorrecto, junto con la del SFN, pueden verse reforzadas por los resultados obtenidos en este trabajo sobre el efecto de las mutaciones en la estabilidad termodinámica (véase más adelante), siempre y cuando se asuma que el SFN es también importante en el mantenimiento de la estructura una vez que la proteína se haya plegado, lo cual no queda explícito en el trabajo de Mirny y Shakhnovich (?).

2.6.4 La estabilidad termodinámica está estrechamente relacionada con la dinámica evolutiva de citocromo b y COX 1

Nuestras evidencias, aportadas en este trabajo de tesis como uno de los objetivos, sugieren que el efecto de las mutaciones sobre la estabilidad es un factor determinante en la dinámica evolutiva de proteínas. Se ha observado que el cambio en estabilidad provocado por mutaciones en residuos enterrados son, en general, más desestabilizadoras que aquellas que afectan a posiciones expuestas, lo cual está en línea con los bajos valores de d_N del interior comparados con los de superficie, independientemente de la proteína estudiada. Además, a pesar del hecho de que el interior de citocromo b evoluciona más

lentamente que su superficie, los residuos enterrados de citocromo b evolucionan de forma semejante a los de la superficie de COX 1, lo que sugiere que pueden encontrarse bajo restricciones evolutivas de igual magnitud (figura 2.7c). Esta observación encaja bien con los resultados de $\Delta\Delta G$, puesto que los obtenidos para el interior de citocromo b son comparables en manitud con los obtenidos para la superficie de COX 1, lo que explica que el centro de citocromo b es más estable frente a los cambios que el interior de COX 1.

Estas evidencias añaden robustez a las teorías del plegado incorrecto y del SFN, expuestas previamente en esta discusión. Con todos estos argumentos en conjunto, se puede concluir que la estabilidad termodinámica es el factor más importante en la dinámica evolutiva de citocromo b y COX 1, no solo a favor del éxito en el plegado de la proteína después de su transcripción, sino también en el mantenimiento de su estructura terciaria.

2.6.5 La influencia de los complejos completos sobre la dinámica evolutiva de sendas subunidades difiere entre citocromo b y COX 1

Como se ha indicado en la introducción a este capítulo, existen evidencias que apoyan que la exposición al solvente de un determinado residuo es determinante en su capacidad evolutiva. Por ello, es razonable especular que una proteína con una baja proporción de residuos expuestos evolucionaría más lentamente en todo su conjunto. Sin embargo, el debate sobre esta hipótesis sigue abierto. Bloom y colaboradores (?) encontraron una correlación negativa entre la tasa evolutiva de la proteína completa y la proporción de residuos expuestos, argumentando que el aumento en la densidad de contactos entre residuos del interior, compensa considerablemente las restricciones provocadas por el decremento de aminoácidos en la superficie, dando lugar a un aumento en la variabilidad global de la secuencia. Aunque esto pueda resultar cierto en determinados conjuntos de proteínas, se observa una alta conservación de COX 1 con respecto a citocromo b, a pesar de que la proporción de residuos expuestos es considerablemente menor. De hecho, Lin et al. (?) detectaron una correlación positiva entre la proporción de residuos expuestos y la tasa de cambios no sinónimos, lo cual sí es coherente con nuestros resultados. Según estos autores, se obvian algunas consideraciones en el trabajo de Bloom et al. (?), como los contactos entre subunidades que forman parte de complejos heteroméricos.

Los complejos citocromo bc₁ y citocromo c oxidasa, a los que pertenecen citocromo b y COX 1, respectivamente, son grandes sistemas heteroméricos con una gran importancia en el metabolismo aeróbico. Las interacciones entre las distintas cadenas de los complejos deben estar implicadas en la capacidad evolutiva de cada proteína, por lo que no se ha obviado en este trabajo. Mientras que los cambios en la estabilidad termodinámica permanecieron distribuidos de igual forma en el caso de citocromo b, tanto como subunidad aislada, como formando parte del complejo III; en el caso de COX 1 se observa una mayor pérdida de estabilidad cuando se realizó la simulación con la proteína formando

2. Dinámica evolutiva de citocromo b y COX 1

parte del complejo IV (figura 2.8). Esto indica, que COX 1 influye en la estabilidad del complejo al que pertenece en mayor medida de lo que lo hace citocromo b, dejando entrever posibles restricciones evolutivas derivadas de la estructura cuaternaria. En el tercer capítulo de esta tesis, entraremos en profundidad en este aspecto.

2.7 Conclusiones

1. Las posiciones invariantes se acumulan en el interior de COX 1, pero pueden estar distribuidas aleatoriamente en toda la proteína en el caso de Citocromo b.
2. El interior de COX 1, al contrario que el interior de Citocromo b, muestra una baja entropía de Shannon, que se aparta significativamente de lo esperado por azar.
3. Las posiciones no constreñidas se distribuyen aleatoriamente en Citocromo b, pero están excluidas selectivamente del interior de COX 1.
4. El gen de COX 1 está sometido a una mayor presión mutacional y a una mayor selección purificadora con respecto a Citocromo b.
5. La estabilidad termodinámica puede explicar las diferencias en el comportamiento evolutivo entre COX 1 y Citocromo b.

Capítulo 3

Evolución de aminoácidos del complejo citocromo c oxidasa (COX) implicados en contactos intermoleculares

3.1 Resumen

Los complejos que forman la cadena respiratoria están codificados por dos genomas (mtDNA y nDNA). Aunque la importancia de la coadaptación intergenómica es bien conocida, las fuerzas y restricciones que gobiernan la coevolución de estas proteínas son desconocidas. Trabajos previos en los que se usa citocromo c oxidasa (COX) como enzima modelo, propusieron la llamada “hipótesis optimizadora”. Según esta teoría, los residuos codificados por mtDNA que se encuentran en contacto con otros codificados por nDNA, evolucionan más deprisa que el resto de las posiciones, favoreciendo la optimización de las interfaces proteína-proteína. En este capítulo, mediante el análisis de datos evolutivos, en combinación con información estructural de COX, dejamos constancia de, que al no discernir correctamente entre los efectos de la interacción entre residuos y otras variables estructurales, se puede llegar a conclusiones engañosas tales como la mencionada “hipótesis optimizadora”. Una vez detectados los factores responsables que adulteraron los trabajos anteriores, llegamos a una conclusión opuesta, que indica que los residuos codificados por mtDNA en contacto con subunidades codificadas por el núcleo están, de hecho, más constreñidos en general que aquellos que no están en contacto. No obstante, los residuos de la superficie de la subunidad 1 (COX 1), que no se encuentran en contacto con ninguna otra subunidad, constituyen una notable excepción, puesto que se encuentran sujetos a una elevada selección purificadora. Además, en este capítulo documentamos que precisamente aquellos residuos codificados por mtDNA, que están

3. Evolución de residuos implicados en contactos intermoleculares

en contacto con otros codificados por el mismo genoma, son aún menos variables que aquellos que interaccionan con polipéptidos codificados por nDNA. Este comportamiento diferencial no pudo ser explicado en términos de estabilidad termodinámica, ya que las interacciones entre subunidades codificadas por mtDNA contribuyen en menor medida a la estabilidad del complejo que aquellas formadas por diferentes genomas. Finalmente, atendiendo a las subunidades codificadas por nDNA, los residuos en contacto están más conservados que aquellos que no están en contacto, con la excepción de la subunidad COX 5A.

3.2 Introducción

Como consecuencia de la importancia de la función de los complejos oligoméricos que forman la cadena respiratoria, aquellas mutaciones que alterasen sus estructuras deberán enfrentarse al consecuente efecto de una elevada presión selectiva. De esta forma, la selección natural favorecerá la coadaptación entre proteínas que interactúan, ya sea hacia la mejora de sus funciones fisiológicas (?) o simplemente hacia la compensación de cambios ligeramente deletéreos que han sido fijados por deriva genética (?). Cualquiera que sea la fuerza evolutiva, existen numerosos trabajos que ponen de manifiesto la importancia biológica de la adaptación intergenómica. Algunos ejemplos muy ilustrativos se hallan en algunos trabajos en los que se llevan a cabo estudios sobre cíbridos xenomitocondriales, que son células con el mtDNA de una especie y el nDNA de otra. Conforme más alejadas filogenéticamente estaban estas especies, más deficiente resultaba la fosforilación oxidativa (?; ?). Estos resultados indicaban la importancia de la coevolución entre genomas presentes en la misma célula. En otros estudios, realizados con copépodos (*Tigriopus californicus*), se llegaron a conclusiones similares. En éstos, se obtuvieron poblaciones en las que los individuos contenían el nDNA de una población y el mtDNA de otra. De forma similar que en el ejemplo anterior, esta población de híbridos mostraba un funcionamiento deficiente en la fosforilación oxidativa (?). Otra línea que evidencia la coevolución entre genomas se expone en estudios de epistasia en los que se observa que mutaciones en mtDNA asociadas a enfermedades en humanos están fijadas u ocurren de forma natural en genomas de mamíferos no humanos (?; ?).

A pesar de que estos estudios enfatizan la relevancia de la coevolución entre proteínas que interaccionan, no arrojan luz sobre las fuerzas y restricciones que modulan tal coevolución. En un intento de explorar la dinámica evolutiva de las regiones que forman las interacciones proteína-proteína, Schmidt y colaboradores (?) llevaron a cabo varios análisis sobre el complejo citocromo c oxidasa (COX), como holoenzima modelo. En este trabajo, se analizó la tasa de sustituciones no sinónimas del conjunto de residuos codificados por mtDNA próximos a aminoácidos codificados por nDNA, y se comparó con el mismo parámetro procedente del conjunto complementario (aquellos residuos codificados por mtDNA que no están en contacto con residuos codificados por nDNA). Sus autores llegaron a la conclusión de que los residuos de origen genómico

3.3. Objetivos

mitocondrial, en contacto con subunidades codificadas en el núcleo, evolucionan más deprisa que aquellos que no están en contacto (?). Estos resultados fueron interpretados como un proceso de optimización de las interacciones entre proteínas. De hecho, los mismos autores denominaron este fenómeno como “interacción optimizadora”. Por otro lado, cuando los residuos de COX codificados por nDNA fueron segregados según su proximidad a subunidades codificadas por mtDNA y se analizó la tasa de sustituciones no sinónimas, la conclusión a la que llegaron fue la opuesta, es decir, aquellos residuos codificados por nDNA en contacto con otros codificados por mtDNA evolucionan más lento que el resto de los residuos del mismo origen genómico (?).

Estos resultados tan llamativos han sido citados frecuentemente en la literatura relacionada (?; ?; ?). Mientras que la lenta evolución de los residuos codificados por nDNA, implicados en contacto con otras subunidades, está en línea con la idea pre establecida de que la evolución de regiones implicadas en interacciones es conservativa, la observación de que las regiones codificadas por mtDNA en contacto con residuos de origen intergenómico evolucionan a mayor velocidad, si se confirmase como un resultado fiable, merece una explicación sólida. Así pues, en este capítulo hemos revisado en profundidad la “hipótesis optimizadora”. Usando un conjunto de especies de gran tamaño muestral y ampliando los análisis estadísticos precedentes, observamos que los residuos codificados por mtDNA en contacto con residuos codificados por nDNA están, en contraste con la “hipótesis optimizadora”, sujetos a mayores restricciones evolutivas que sus residuos homólogos que no están en contacto. Además, exponemos detalladamente por qué los autores anteriores fallaron en la interpretación de sus resultados. Adicionalmente, mostramos una intrigante diferencia en las tasas evolutivas de residuos codificados por mtDNA en contacto con residuos de origen intra- (mtDNA) e intergenómico (nDNA).

3.3 Objetivos

En el trabajo de investigación que se desarrolla en este último capítulo, se han combinado las conclusiones alcanzadas en el capítulo anterior sobre la evolución diferencial entre interior y superficie de proteínas, con el conocimiento actual sobre el efecto de las interacciones interproteicas en la dinámica evolutiva de proteínas mitocondriales. De forma sucinta, los objetivos planteados fueron:

1. Desarrollar un modelo que permitiera realizar análisis evolutivos de regiones de la estructura cuaternaria de proteínas, acotadas según su exposición al solvente y la distancia a residuos pertenecientes a otros péptidos que forman parte del mismo complejo.
2. Describir la dinámica evolutiva de residuos del complejo citocromo c oxidasa que se encuentran implicados en contactos intermoleculares con otras subunidades, atendiendo a su origen genómico.

3. Evolución de residuos implicados en contactos intermoleculares

3.4 Material y métodos

3.4.1 Datos

Los datos analizados en este capítulo son referentes a las secuencias y estructuras de las 13 subunidades distintas que forman el complejo citocromo c oxidasa (complejo IV). La información genética fue obtenida del NCBI (*National Center for Biotechnology Information*, <http://www.ncbi.nlm.nih.gov>). Ésta comprende, por un lado, las secuencias de los genomas mitocondriales de 371 especies de mamíferos y por otro, entre 14 y 30 secuencias ortólogas de los genes nucleares que codifican para las 10 subunidades del complejo (véase apéndice F). Para cada subunidad, se llevó a cabo un alineamiento múltiple de secuencias mediante ClustalW. Los alineamientos resultantes están disponibles en <http://mecom.hval.es/datasets>.

La información estructural del complejo IV fue obtenida del archivo identificado como 2occ (citocromo c oxidasa de *Bos taurus*) en la base de datos PDB (*Protein Data Bank*, <http://www.pdb.org/>).

La edad de cada subunidad fue estimada mediante el servidor *ProteinHistorian*, cuya interfaz de usuario está disponible vía web en <http://lighthouse.ucsf.edu/ProteinHistorian/> (?).

3.4.2 Diseño e implementación del modelo de estudio

La función principal que desempeña el algoritmo diseñado para el análisis evolutivo llevado a cabo en este trabajo, es separar los alineamientos de las distintas subunidades en otros nuevos en base a las características estructurales de los aminoácidos para los que codifican, concretamente, la exposición del residuo al solvente y su proximidad a subunidades distintas a las que pertenecen. En la figura 3.1 se representa un esquema del algoritmo implementado en este trabajo, tomando como ejemplo COX 1, una subunidad codificada por mtDNA. El esquema está dividido en 4 fases: (i) Datos, (ii) Segregación de alineamientos, (iii) Análisis y (iv) Contraste. La primera fase consiste en la recolección de datos, el alineamiento de las secuencias y la caracterización estructural de los tripletes (véase apartado 3.4.3). La segunda fase, representada en el esquema con un color más intenso, es la división de los alineamientos en otros, compuestos por subconjuntos de codones procedentes del alineamiento principal. Los criterios de clasificación se exponen en el recuadro inferior del mismo esquema. La tercera fase, consiste en la caracterización evolutiva de los alineamientos (véase apartado 3.4.4) y la cuarta, implementa una prueba estadística Z destinada a evaluar si los valores de R son significativamente distintos de cero.

Este proceso ha sido implementado en un software desarrollado *ad hoc* para este trabajo, al que hemos denominado con las siglas MECOM (*Molecular Evolution of*

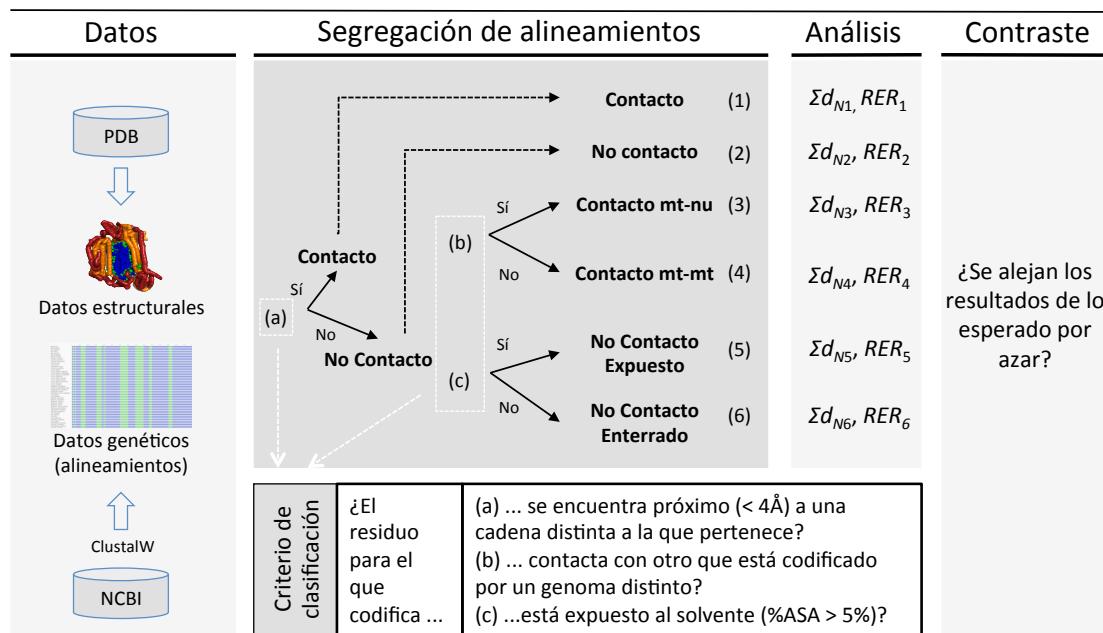


Figura 3.1: Esquema del modelo de análisis. El flujo de trabajo se divide en cuatro fases, de izquierda a derecha: (i) Datos, (ii) Segregación de alineamientos, (iii) Análisis y (iv) Contraste. Los detalles de cada una de las fases se explican en el texto.

protein COMplexes). Éste, junto con un manual de uso, ha sido almacenado en repositorios de acceso público (CPAN, <http://search.cpan.org/~hvalverde/Mecom-1.15/> y GitHub <https://github.com/hvpareja/Mecom>) y se ha desarrollado un sitio web dedicado al desarrollo y mejora del software (<http://mecom.hval.es/>). En el apéndice A se describen los detalles técnicos del programa.

3.4.3 Caracterización estructural de los residuos

A partir del modelo atómico del complejo citocromo c oxidasa, se llevó a cabo un análisis estructural atendiendo a la superficie expuesta al solvente de cada residuo y la proximidad a otros residuos pertenecientes a otras cadenas del complejo. Para el primer caso, se procedió, de forma similar a la descrita en la sección 2.4.2 del capítulo 2, el porcentaje de superficie expuesta al solvente de cada residuo (%ASA_i) mediante el programa DSSP (*Dictionary of Protein Secondary Structure*). Se consideró que el residuo, *i*, está enterrado si %ASA_i < 5 %. Por otra parte, el cálculo de las distancias euclídeas entre aminoácidos se realizó mediante software desarrollado para este trabajo. Primero, se obtuvo la distancia en angstroms para cada par de grupos atómicos de cada par de residuos pertenecientes a cadenas distintas, en segundo lugar, se determinó que la distancia entre dos residuos se corresponde con la distancia más corta entre el par de grupos atómicos más cercanos (figura 3.2). Se determinaron como residuos en contacto aquellos pares

3. Evolución de residuos implicados en contactos intermoleculares

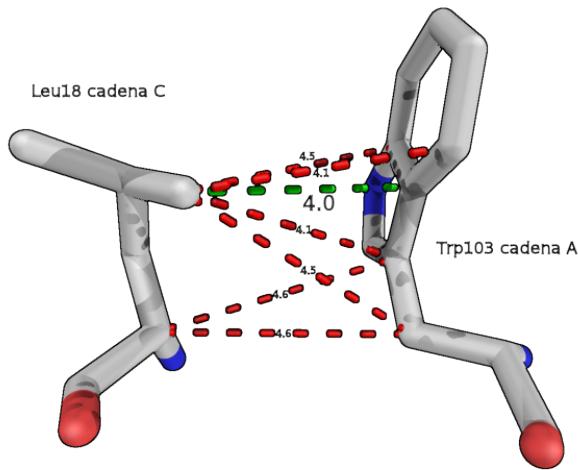


Figura 3.2: Determinación de distancia entre dos residuos. Se calculó la distancia eculídea entre cada par de coordenadas pertenecientes a grupos atómicos de dos residuos localizados en cadenas distintas. Posteriormente, se determinó la distancia atómica más corta como la proximidad entre los dos aminoácidos. En el caso del ejemplo representado en la imagen, la distancia entre Leu18 de la cadena C (COX 3) y Trp103 de la cadena A (COX 1) es de 4.0 Å.

cuya distancia resultó inferior o igual a 4 Å.

3.4.4 Caracterización evolutiva de la subunidades

Tasa de sustituciones no sinónimas

A partir de cada alineamiento i , numerados del 1 al 5 en el ejemplo de la figura 3.1, se calcularon para cada par de secuencias, j , las tasas de sustituciones no sinónimas por sitio no sinónimo (d_{Nj}) mediante el método de Nei & Gobori (?) implementado en PAML 4.6 (?; ?). Se calculó entonces el sumatorio de los valores obtenidos para el total los pares de secuencias p , tal y como se expresa en la ecuación 3.1.

$$\Sigma d_{Ni} = \sum_{j=1}^p d_{Nj} \quad (3.1)$$

Además, para el caso de las subunidades codificadas por mtDNA (COX 1, COX 2 y COX 3), las cuales son las cadenas con mayor número de aminoácidos del complejo IV (514, 227 y 261, respectivamente), fue posible generar una distribución aleatoria de Σd_{Ni} . Para ello, a partir del alineamiento original de cada subunidad, se tomaron al azar un número de nucleótidos igual al del alineamiento correspondiente al subconjunto de residuos i . Posteriormente se calculó, para este conjunto aleatorio, el valor de Σd_{Ni}

3.4. Material y métodos

(ecuación 3.1). Este proceso se repitió 10^4 veces, obteniendo una distribución aleatoria empírica de esta variable.

Con el ánimo de comprobar si el comportamiento de los valores de d_N , calculados para la muestra completa de secuencias, es uniforme a lo largo de la escala filogenética, se llevó a cabo un análisis basado en linajes mediante un método de máxima verosimilitud¹ (modelo F3x4), implementado en el programa `codeml` del paquete PAML (?). Para este propósito, se obtuvo, en primer lugar, el árbol filogenético de siete especies de mamíferos (*Bos taurus*, *Sus scrofa*, *Equus caballus*, *Ailuropoda melanoleuca*, *Mus musculus*, *Rattus norvegicus* y *Cavia porcellus*) para las que se disponía de las secuencias de los genes que codifican para las 13 subunidades de COX, tomando como grupo externo la especie *Danio rerio*.

Tasa de interacción

Dado que el interés principal de este trabajo es la caracterización de las diferencias en el comportamiento evolutivo de distintos subconjuntos de residuos de cada subunidad del complejo, se calculó la denominada *tasa de interacción*, descrita previamente en la literatura (?). Este valor, representado por $R_{1,2}$, es el cociente entre el valor de Σd_N del conjunto 1 y el conjunto 2 (ecuación 3.2).

$$R_{1,2} = \frac{\Sigma d_{N1}}{\Sigma d_{N2}} \quad (3.2)$$

De esta forma, si el resultado de la tasa de interacción para un par de conjuntos es significativamente superior a 1 ($R_{1,2} > 1$), puede interpretarse que el conjunto 1 presenta una tasa evolutiva significativamente superior a la del conjunto 2, y viceversa. Por el contrario, si el valor de $R_{1,2}$ no es distinto de 1, se puede deducir que las tasas evolutivas de los dos conjuntos no difieren. Para contrastar estadísticamente si cada valor de $R_{i,j}$ fue significativamente distinto de 1, se llevó a cabo una prueba estadística Z.

Tasa de evolución relativa (RER)

Con el objetivo de añadir robustez a los resultados obtenidos en este trabajo, se calculó otro parámetro evolutivo denominado *tasa de evolución relativa* (RER). Para este propósito se usó el programa MEGA 5 (?), que implementa un método de máxima verosimilitud y hace uso del modelo de sustitución de Jones-Taylor-Thornton (?) para estimar dicho parámetro. Se calcularon las diferencias en las tasas evolutivas entre los sitios, usando como modelo una distribución Gamma (+G) discreta. Se consideraron 5 categorías Gamma. Los resultados obtenidos fueron escalados de forma que el valor

¹«Maximun likelihood»

3. Evolución de residuos implicados en contactos intermoleculares

medio de las tasas de todos los sitios fuese 1. Así, los sitios con valores menores a 1 evolucionarían más lento que la media y los sitios con valores mayores a 1 evolucionarían más deprisa.

3.4.5 Cambios en la estabilidad termodinámica ($\Delta\Delta G$)

Al igual que en el trabajo recogido en el capítulo 2 (página 38), se hizo uso del programa FoldX versión 3.0 (?) para calcular los cambios en la estabilidad termodinámica $\Delta\Delta G$, provocados por una mutación puntual en cada una de las posiciones de cada subunidad (comando `alascan`). Previamente, y al igual que en el capítulo anterior, se llevó a cabo una optimización del modelo (comando `RepairPDB`). Posteriormente, se calculó la media de $\Delta\Delta G$ de los conjuntos de residuos pertenecientes a las categorías “No Contacto Expuesto”, “Contacto mt-mt” y “Contacto mt-nu”.

3.5 Resultados

En el capítulo anterior, se ha descrito la dinámica evolutiva de dos proteínas que forman parte de complejos multiprotéicos de la membrana mitocondrial interna. Los resultados obtenidos sobre COX 1, perteneciente al complejo IV o citocromo c oxidasa, nos llamaron especialmente la atención por la alta conservación de sus residuos y la susceptibilidad de su estructura ante mutaciones. Además, un cambio en cualquier residuo de COX 1 tuvo una repercusión significativamente mayor sobre la estabilidad del complejo al completo que sobre la proteína en sí misma (figura 2.8). Motivados por estas observaciones, se quiso abordar un nuevo estudio en el que se pusiera de manifiesto el comportamiento evolutivo de los contactos intermoleculares existentes entre COX 1 y el resto de las cadenas del complejo IV.

3.5.1 Comparativa entre los dos modelos de estudio

En el trabajo de Schmidt y colaboradores (?), se llevaron a cabo análisis de las tasas de sustitución de residuos en contacto con otros de distinta cadena del complejo. A partir de un alineamiento de secuencias de 26 especies de mamíferos distintas, se construyeron dos alineamientos nuevos: el primero, denominado $ABC_{Contacto\ mt-nu}$, estaba formado por nucleótidos del genoma mitocondrial (cadenas A, B y C) que codificaban para residuos en contacto con cadenas codificadas por el genoma nuclear (cadenas D-M). El segundo estaba formado por el conjunto complementario, $(ABC_{Contacto\ mt-nu})^c$, es decir, por aquellos nucleótidos que codificaban para residuos que no estaban en contacto con cadenas codificadas por nDNA. Esta división de posiciones se representa en el recuadro (a) de la figura 3.3. Con estos datos, se calculó la tasa de interacción, $R_{ABC_{Contacto\ mt-nu}, ABC_{(Contacto\ mt-nu)^c}}$, cuyo resultado fue significativamente mayor a la

unidad (véase apartado 3.4.4 de Material y Métodos). Esto indicaba, según la interpretación propuesta por los mismos autores, que aquellos aminoácidos codificados por mtDNA, cercanos en la estructura cuaternaria a otros codificados por nDNA, presentan una tasa evolutiva más alta que el resto de los residuos del mismo origen genómico.

Como punto de partida para los análisis realizados en este capítulo, se reprodujeron los resultados del trabajo de Schmidt y colaboradores con un mayor número de secuencias ortólogas en los alineamientos originales (371 frente a 26). El resultado de la tasa de interacción de nuestro ensayo fue de $1.805 \pm 4 \times 10^{-4}$, significativamente mayor a 1. Aunque este resultado es coherente con el trabajo anterior, pensamos que el modelo experimental del trabajo de Schmidt et al., fue poco minucioso en la selección de los conjuntos de residuos llevados a estudio, especialmente el en el caso del grupo ABC_{Contacto-mt-nu}^c, que incluye una colección de residuos muy heterogénea, lo que les llevó a obviar dos aspectos muy importantes desde el punto de vista estructural: (i) la exposición al solvente, que como se ha demostrado en el trabajo recogido en el capítulo anterior (página 54), es influyente en la dinámica evolutiva y (ii) la interacción entre residuos de distintas cadenas, pero codificadas por el mismo genoma mitocondrial (interacciones entre COX 1, COX 2 y COX 3). Teniendo en cuenta estos aspectos, en nuestro modelo se incluyó, por un lado, el alineamiento ABC_{Contacto mt-nu}, y por otro, dos alineamientos construidos a partir del grupo (ABC_{Contacto mt-nu})^c: uno con los residuos en contacto con cadenas codificadas por el genoma mitocondrial, ABC_{Contacto mt-mt}, y otro con aquellos residuos libres² y expuestos al solvente (%ASA ≥ 5%), ABC_{No Contacto Expuesto} (figura 3.3b). Además, se calcularon las tasas de interacción a partir de los alineamientos de cada cadena por separado. Así, los resultados mostrados en la tabla 3.4, aportaron más información sobre la contribución de cada subunidad codificada por mtDNA a los resultados obtenidos con los alineamientos de las tres proteínas concatenados.

3.5.2 Comportamiento evolutivo de las distintas regiones de las cadenas codificadas por mtDNA

Una vez que segregados los residuos en sus correspondientes categorías, atendiendo a las propiedades estructurales mencionadas anteriormente, y habiendo reproducido los resultados de los análisis precedentes (?), quisimos abordar el estudio de la dinámica evolutiva de cada categoría de una forma más robusta. Para ello, tomamos diversas aproximaciones metodológicas. Una de ellas consiste en el cálculo de la denominada tasa de evolución relativa (*RER*). Este método, implementado en el programa MEGA 5 (?), permite estimar las diferencias en la tasa evolutiva entre distintos sitios de una misma secuencia. En este parámetro, el valor de la unidad representa la tasa evolutiva media del conjunto completo, así, si el valor de *RER* es superior a 1 en un subconjunto, indica que la tasa evolutiva de éste es superior a la media y, por el contrario, un valor inferior a 1, indica que la tasa evolutiva es inferior a la media. En la tabla 3.1, se observa que

²Entiéndase «residuo libre» como aquel aminoácido que no está en contacto con otra subunidad del complejo, es decir, perteneciente subconjunto denominado “No Contacto”.

3. Evolución de residuos implicados en contactos intermoleculares

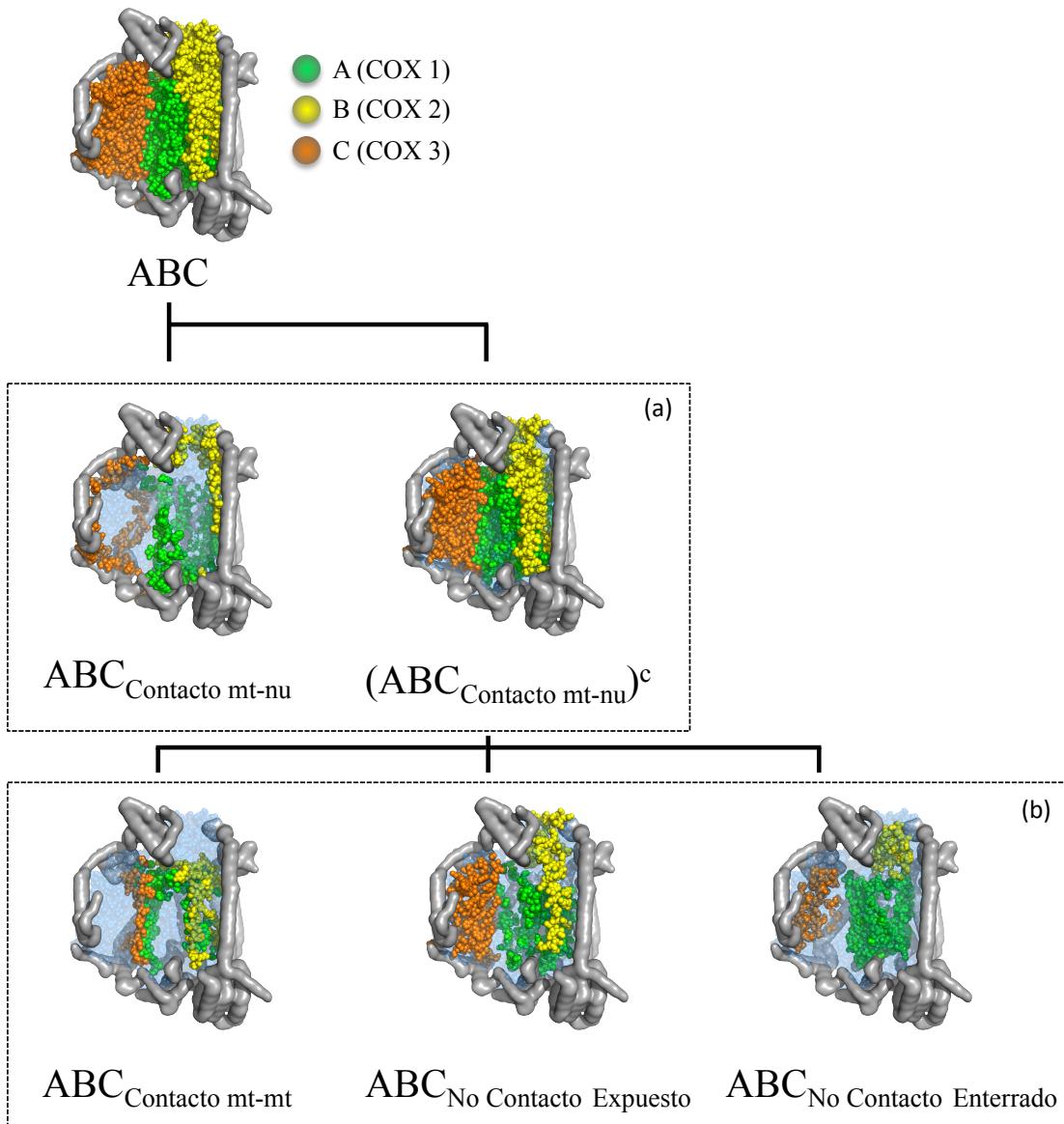


Figura 3.3: Estructura de los subconjuntos de residuos de COX codificados por el genoma mitocondrial. Las subunidades codificadas por nDNA se representan como superficie en color gris, mientras que las subunidades codificadas por mtDNA se representan como esferas de distinto color. En el recuadro (a) se representan los subconjuntos de posiciones construidos en el modelo de Schmidt (?), y en el recuadro (b) se representan los subconjuntos adicionales construidos en el modelo que se propone en este trabajo.

3.5. Resultados

el grupo de residuos pertenecientes a las categorías “Contacto mt-mt” y “No Contacto Enterrado” son los más constreñidos independientemente de la subunidad considerada, ya que muestran valores de *RER* inferiores a la unidad.

Tabla 3.1: Tasas de evolución relativa (RER)

	RER			
	ABC	A	B	C
No Contacto Expuesto	1.27 ± 2.26	1.03 ± 2.17	1.52 ± 1.87	1.48 ± 2.54
Contacto mt-mt	0.41 ± 0.92	0.26 ± 0.31	0.32 ± 0.45	0.88 ± 1.82
Contacto mt-nu	1.60 ± 2.33	2.33 ± 2.98	1.21 ± 1.67	0.90 ± 1.26
No Contacto Enterrado	0.41 ± 0.62	0.39 ± 0.53	0.78 ± 1.00	0.16 ± 0.22

Los valores de la tasas de evolución relativa que se muestran en la tabla se expresan como media ± desviación típica. Los valores inferiores a 1 indican que la tasa evolutiva del conjunto correspondiente es menor que la media y, por el contrario, los valores superiores a 1 indican que la tasa evolutiva es mayor que la media. Los valores que no son distintos de 1 se corresponden con los residuos cuya tasa evolutiva no difiere de la media.

Con el fin de fortalecer las conclusiones obtenidas a partir de estos resultados, se han representado en la tabla 3.2 los valores de Σd_N de cada una de las categorías. Este valor es el resultado del sumatorio de todas las tasas de sustituciones no sinónimas por sitio no sinónimo (d_N) de todos los pares de secuencias de la muestra (véase apartado 3.4.4 de Material y Métodos para más detalles). Estos resultados están en línea con los valores de *RER*, sin embargo, ambos parámetros tan solo ofrecen una visión global de la capacidad evolutiva de cada subunidad, pero no ofrecen información suficiente para determinar si esta dinámica evolutiva está realmente presente en la mayoría de los linajes de mamíferos. Para abordar esta cuestión, se realizó un nuevo análisis basado en la filogenia. Para ello, se obtuvo el valor de d_N de cada rama del árbol filogenético construido con un conjunto de 7 especies de mamíferos (tabla 3.3). Todos los linajes sin excepción mostraron un valor mayor de d_N en el caso de los residuos codificados por mtDNA en contacto con subunidades codificadas por nDNA, en comparación con los residuos que contactan con subunidades codificadas por mtDNA.

Estos resultados, en conjunto, sugieren que aquellos residuos codificados por mtDNA, en contacto con residuos codificados por el mismo genoma, están sujetos a mayores restricciones que aquellos implicados en interacciones con subunidades codificadas por nDNA. Además, este comportamiento parece estar presente en la mayoría de los linajes de mamíferos (Tabla 3.3).

3. Evolución de residuos implicados en contactos intermoleculares

Tabla 3.2: Tasas de sustituciones no sinónimas (Σd_N)

	Σd_N			
	ABC	A	B	C
No Contacto Expuesto	5954 ± 10	2901 ± 9	10605 ± 104	8404 ± 50
Contacto mt-mt	2508 ± 7	939 ± 5	3034 ± 27	5820 ± 83
Contacto mt-nu	7050 ± 15	7603 ± 40	7129 ± 56	6450 ± 44
No Contacto Enterrado	1968 ± 3	1252 ± 4	5586 ± 70	1230 ± 14

Cada valor de esta tabla fue calculado como el sumatorio de los valores de las tasas de sustituciones no sinónimas por sitio no sinónimo (d_N) de cada par de secuencias de cada alineamiento. Los datos se expresan como media ± desviación típica.

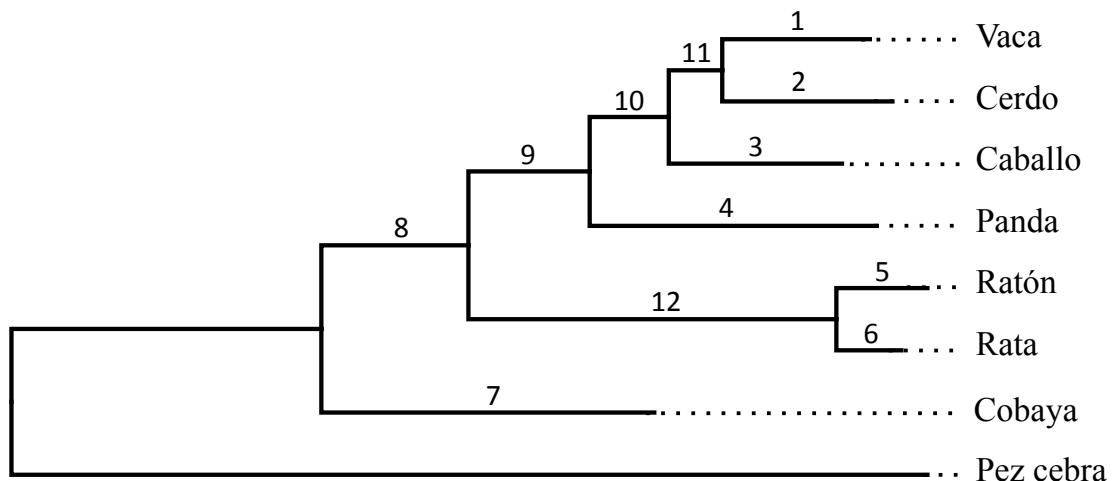
3.5.3 Tasas de interacción de subunidades codificadas por mtDNA

A partir de los valores de Σd_N , obtenidos para cada categoría de residuos (tabla 3.2), se calcularon los valores de las tasas de interacción (tabla 3.4). Éstos indicaron que, en el caso de las cadenas B (COX 2) y C (COX 3), los aminoácidos en contacto con otros codificados por nDNA, están más constreñidos que los residuos libres y expuestos. Sin embargo, para la cadena A (COX 1) se obtuvo un valor de $R_{A_{\text{Contacto mt-nu}}, A_{\text{No Contacto Expuesto}}}$ de 2.3 ± 10^{-3} . Esto indica que el hecho de que el resultado de $R_{ABC_{\text{Contacto mt-nu}}, ABC_{\text{No Contacto Expuesto}}}$ sea mayor que uno, se debe en gran medida al excepcional comportamiento evolutivo de COX 1. Por otro lado, al comparar las tasas evolutivas de las interacciones intragenómicas (entre residuos codificados por el mismo genoma) e intergenómicas (entre residuos codificados por distinto genoma), se observa que los aminoácidos codificados por mtDNA que están en contacto con otros codificados también por mtDNA, se encuentran más constreñidos que aquellos que interactúan con subunidades codificadas por nDNA, salvo en COX 3, cuyo valor de $R_{C_{\text{Contacto mt-nu}}, C_{\text{Contacto mt-mt}}}$, resultó estar en torno a la unidad.

Como se observa en la tabla 3.4, el valor de $R_{ABC_{\text{Contacto mt-nu}}, ABC_{\text{No Contacto Expuesto}}}$ es notablemente menor que el valor de $R_{ABC_{\text{Contacto mt-nu}}, (ABC_{\text{Contacto mt-nu}})^c}$, obtenido mediante el modelo de Schmidt ($1.1 \pm 3 \times 10^{-4}$ y $1.8 \pm 4 \times 10^{-4}$, respectivamente), dejando en evidencia el efecto de la exclusión de los residuos enterrados y los implicados en interacciones con cadenas de origen genético mitocondrial.

3.5. Resultados

Tabla 3.3: Patrón de sustituciones no sinónimas en las subunidades del complejo IV en distintos linajes de mamíferos.



Rama	mtDNA				nDNA	
	Contacto mt-mt	Contacto mt-nu	Contacto	No Contacto Expuesto	Contacto	No Contacto Expuesto
1	0.006	0.033	0.026	0.076	0.014	0.044
2	0.006	0.046	0.035	0.037	0.011	0.030
3	0.003	0.029	0.021	0.034	0.014	0.028
4	0.006	0.046	0.033	0.090	0.021	0.028
5	0.005	0.039	0.029	0.036	0.011	0.015
6	0.003	0.020	0.015	0.034	0.012	0.389
7	0.005	0.036	0.029	0.072	0.017	0.046
8	0.005	0.038	0.027	0.068	0.017	0.035
9	0.000	0.016	0.008	0.058	0.011	0.032
10	0.003	0.020	0.011	0.000	0.002	0.005
11	0.006	0.008	0.011	0.000	0.010	0.021
12	0.006	0.062	0.044	0.119	0.034	0.064

Se construyó el árbol filogenético para 7 especies de mamíferos para las que se disponía de las secuencias genéticas de los 13 genes que codifican para el complejo citocromo c oxidasa (*Bos taurus*, *Sus scrofa*, *Equus caballus*, *Ailuropoda melanoleuca*, *Mus musculus*, *Rattus norvegicus* y *Cavia porcellus*), usando *Danio rerio* como grupo externo. En la tabla, se muestra el número de sustituciones no sinónimas por sitio no sinónimo para cada uno de los grupos de residuos indicados en la cabecera.

3. Evolución de residuos implicados en contactos intermoleculares

Tabla 3.4: Tasas de interacción de las subunidades codificadas por mtDNA.

	$R_{1,2}^{\frac{1}{2}}$			
	ABC	A	B	C
<i>Contacto mt-nu</i>				
<i>No Contacto Expuesto</i>	$1.1 \pm 3 \times 10^{-4}$	2.3 ± 10^{-3}	$0.7 \pm 5 \times 10^{-4}$	$0.7 \pm 5 \times 10^{-4}$
<i>Contacto mt-mt</i>				
<i>No Contacto Expuesto</i>	$0.4 \pm 3 \times 10^{-4}$	$0.3 \pm 6 \times 10^{-4}$	$0.3 \pm 3 \times 10^{-4}$	0.7 ± 10^{-4}
<i>Contacto</i>				
<i>No Contacto Expuesto</i>	$0.8 \pm 2 \times 10^{-4}$	$1.4 \pm 5 \times 10^{-4}$	$0.5 \pm 3 \times 10^{-4}$	$0.7 \pm 4 \times 10^{-4}$
<i>Contacto mt-nu</i>	$2.6 \pm 2 \times 10^{-3}$	7.2 ± 0.014	$2.4 \pm 2 \times 10^{-3}$	1.0 ± 10^{-3}
<i>Contacto mt-mt</i>				

[§] La tasa de interacción, $R_{1,2}$, es el cociente $\frac{\sum d_{N_1}}{\sum d_{N_2}}$, siendo los grupos 1 y 2 los indicados en el numerador y el denominador, respectivamente, en las fracciones de la primera columna.

3.5.4 La estabilidad termodinámica no explica las diferencias entre el comportamiento evolutivo de los residuos en contacto con cadenas mitocondriales y nucleares

Dadas las diferencias observadas en el comportamiento evolutivo de los residuos que contactan con subunidades codificadas por mtDNA, y los que contactan con subunidades codificadas por nDNA, podrían existir diferencias en la contribución de estos conjuntos a la estabilidad termodinámica del complejo, de modo que pudieran explicar este comportamiento. Para ensayar esta hipótesis, se realizó un *alascan* de la misma forma en la que se procedió en el capítulo anterior (apartado 2.4.6) y se analizaron estadísticamente los resultados de $\Delta\Delta G$ correspondientes a los residuos pertenecientes a los distintos grupos (tabla 3.5). Los cambios menos desestabilizadores fueron los del grupo de los residuos “No Contacto Expuesto”. Por otra parte, los residuos “Contacto mt-nu” fueron mucho menos tolerantes a los cambios que cualquier otra categoría. Por lo tanto, el alto grado de conservación observado en los residuos “Contacto mt-mt”, no está explicado por la estabilidad termodinámica, lo que sugiere que hay otros aspectos estructurales o funcionales que deben explicar estas diferencias.

3.5.5 Los residuos expuestos y libres de COX 1 se encuentran excepcionalmente conservados

Según las tasas de interacción mostradas en la tabla 3.5, los residuos en contacto están, en general, más conservados que aquellos que están expuestos y libres. Sin embargo, COX 1 es la excepción en estos resultados. Mientras que para las subunidades COX 2 y COX 3, el valor de la tasa de interacción es significativamente menor que 1, para COX 1 es significativamente mayor a 1 ($1.4 \pm 5 \times 10^{-4}$). Este resultado puede ser interpretado de dos formas: por un lado, puede ser debido a un valor alto de la tasa de sustituciones no sinónimas de los residuos en contacto (el numerador en la ecuación 3.2), lo cual sería

3.5. Resultados

Tabla 3.5: Cambios en la estabilidad termodinámica.

	$\Delta\Delta G$ (kJ/mol)			
	ABC	A	B	C
No Contacto Expuesto	1.41 ± 1.51 (†,§§)	1.64 ± 1.69 (†)	1.20 ± 1.09 (†,§§)	1.16 ± 1.38 (§)
Contacto mt-mt	1.78 ± 1.97 (*)	1.98 ± 1.65	1.92 ± 2.61	1.06 ± 1.52 (§)
Contacto mt-nu	1.92 ± 1.60 (**)	2.03 ± 1.55 (*)	1.99 ± 1.47 (**)	1.69 ± 1.73 (*,†)

Cada dato se expresa como media ± desviación típica

El número de residuos “No Contacto Expuesto” de las cadenas A, B y C fue de 144, 56 y 92, respectivamente.

El número de residuos “Contacto mt-mt” de las cadenas A, B y C fue de 95, 54 y 37, respectivamente.

El número de residuos “Contacto mt-nu” de las cadenas A, B y C fue de 120, 74 y 82, respectivamente.

* Significativamente distinto de “No Contacto Expuesto”, (*) $p < 0.05$, (**) $p < 0.0005$.

† Significativamente distinto de “Contacto mt-mt”, (†) $p < 0.05$, (††) $p < 0.0005$.

§ Significativamente distinto de “Contacto mt-nu”, (§) $p < 0.05$, (§§) $p < 0.0005$.

un argumento a favor de la “hipótesis optimizadora” de Schmidt, y por el otro, este valor puede ser debido a un valor muy bajo en la tasa de sustituciones no sinónimas de los residuos libres y expuestos (el denominador). Para abordar esta cuestión, se representaron gráficamente los valores de Σd_N de cada conjunto de cada subunidad (figura 3.4). Se observa que los cambios en COX 1 están mucho más constreñidos que en el resto de las subunidades, sobretodo en el caso de los aminoácidos expuestos y libres. Estos resultados indican claramente que las tasas de interacción mayores que 1 en el caso de COX 1, se deben a la excepcional conservación de los residuos libres y expuestos, y no ofrecen evidencias de un proceso evolutivo de optimización de contactos.

La simple comparación de los valores de Σd_N de los conjuntos “No Contacto Expuesto” entre las tres subunidades COX 1, COX 2 y COX 3, mostrada en la figura 3.4, deja una evidencia sobre la particular conservación de este grupo de residuos de COX 1. No obstante, cabe preguntarse si este conjunto está significativamente más conservado que el resto de los residuos de la misma subunidad o es, simplemente, un reflejo de la baja variabilidad de COX 1 en su totalidad. Para abordar esta cuestión, realizamos un ensayo

3. Evolución de residuos implicados en contactos intermoleculares

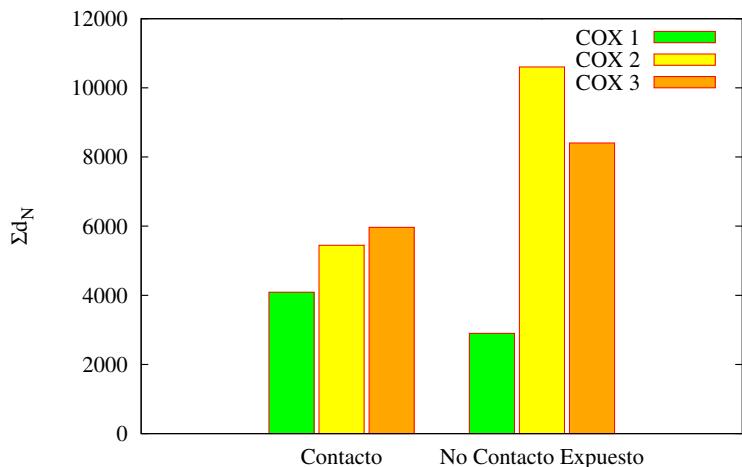


Figura 3.4: Los residuos libres y expuestos de COX 1 están excepcionalmente conservados. En la figura se representa el valor de la tasa de sustituciones sinónimas (Σd_N) de cada uno de los subconjuntos de residuos construidos a partir de las cadenas codificadas por mtDNA. El valor que toma el subconjunto “No Contacto Expuesto” de COX 1 está muy por debajo del observado en las otras dos cadenas procedentes del genoma mitocondrial.

basado en un muestreo aleatorio³. Esto es, para cada cadena codificada por mtDNA, se seleccionaron aleatoriamente del correspondiente alineamiento múltiple, un número de nucleótidos al azar igual al número de nucleótidos que codifican para los residuos pertenecientes al grupo de “No Contacto Expuesto”. Posteriormente, se calculó el valor de Σd_N para este conjunto aleatorio. Este proceso se repitió 10^4 veces por cada subunidad con el fin de construir una distribución aleatoria empírica, la cual se usó para contrastar los valores de Σd_N del conjunto de residuos libres y expuestos. Tal y como se observa en la figura 3.5, los residuos “No Contacto Expuesto” de las subunidades COX 2 y COX 3 son los más variables ($p = 0.003$ y 0.045 , respectivamente), mientras que ocurre lo contrario en los residuos de “No Contacto Expuesto” de COX 1, los cuales representan el grupo de residuos más conservados de esta subunidad ($p = 0.064$).

3.5.6 Existe correlación entre $\Delta\Delta G$ y Σd_N

Con el fin de aislar las posibles fuerzas que gobiernan la excepcional conservación del grupo de residuos expuestos y libres de COX 1, se llevó a cabo una prueba de correlación entre los resultados obtenidos en las simulaciones de mutagénesis con FoldX (**alascan**), y los valores de Σd_N para cada uno de los grupos “Contacto” y “No Contacto Expuesto” de cada subunidad codificada por mtDNA. La gráfica de dispersión Σd_N versus $\Delta\Delta G$ se muestra en la figura 3.6. Se obtuvo una correlación negativa estadísticamente signifi-

³La expresión «muestreo aleatorio» es una traducción adaptada del término estadístico «bootstrapping».

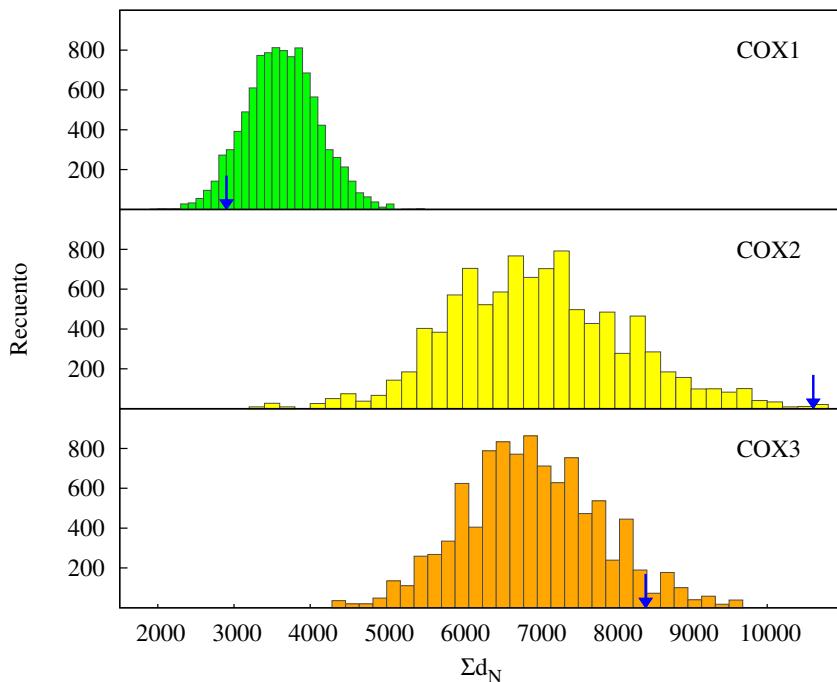


Figura 3.5: El comportamiento evolutivo del conjunto de residuos “No Contacto Expuesto” de COX 1 difiere del mostrado por COX 2 y COX 3. Se construyó una distribución aleatoria empírica para cada subunidad (véase texto para más detalles) y se usaron para contrastar el valor de Σd_N del conjunto de residuos libres y expuestos de cada cadena codificada por mtDNA, los cuales se muestran mediante flechas azules.

3. Evolución de residuos implicados en contactos intermoleculares

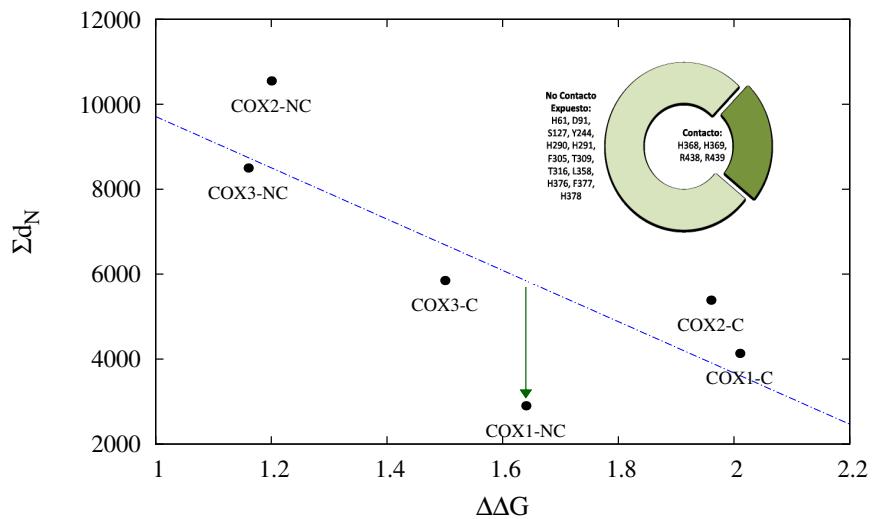


Figura 3.6: Los residuos libres y expuestos de COX 1 están mucho más conservados de lo esperado por el potencial efecto desestabilizador en el complejo. Para cada subunidad codificada por mtDNA, se tomaron los valores de Σd_N de los dos subconjuntos de residuos “Contacto” (C) y “No Contacto Expuesto” (NC) y se representaron frente a los valores de $\Delta\Delta G$ de los mismos subconjuntos. Se obtuvo una correlación negativa ($n = 6$, $r = -0.774$, $p = 0.035$) que mejoraba considerablemente cuando los datos correspondientes a COX 1-NC se excluían del análisis ($n = 5$, $r = -0.890$, $p = 0.021$). El diagrama de sectores, inserto en la gráfica, representa la proporción de residuos funcionales descritos en los dos subconjuntos procedentes de COX 1.

cativa entre estas dos variables ($p < 0.035$). Estos resultados son coherentes con los del capítulo 2, en el que ya se observó que los residuos más desestabilizadores son también los que están más constreñidos. No obstante, no se puede afirmar que el alto grado de conservación de los residuos “No Contacto Expuesto” de COX 1 sea debido a su implicación en la estabilidad del complejo, pues tal y como se observa en la figura 3.6, el valor correspondiente es un punto díscolo, con un valor de Σd_N mucho menor de lo esperado.

3.5.7 Relevancia de los residuos funcionales de COX 1 en su dinámica evolutiva

Después de realizar una descripción de los parámetros estructurales de COX 1, se han descartado posibles explicaciones del comportamiento evolutivo anómalo de los residuos expuestos y libres pertenecientes a esta subunidad. Seguidamente, basándonos en que, generalmente, los residuos funcionales están especialmente conservados en la evolución (?), planteamos la hipótesis de que los residuos funcionales de COX 1 podrían estar más representados en aquel conjunto que presenta una baja tasa de cambios no sinónimos, es decir, el grupo de los residuos libres y expuestos.

El diagrama de sectores incluido en la figura 3.6 muestra la proporción en cada grupo de todos los residuos descritos en *UniProt* (<http://www.uniprot.org/>) como “importantes en la función” de COX 1. Como ejemplos de estos residuos están los aminoácidos implicados en la unión con el grupo hemo, Cu⁺⁺ o Mg⁺⁺, el bombeo de protones y la transferencia de electrones (?; ?). Se observa que la mayoría (76.4 %) pertenecen al grupo de “No Contacto Expuesto”.

3.5.8 Tasas de interacción de las cadenas codificadas por nDNA

Llevando a cabo el mismo proceso que con las cadenas codificadas por mtDNA, se obtuvieron las tasas de interacción $R_{1,2}$ entre los residuos en contacto y los expuestos y libres de las cadenas codificadas por nDNA. En la figura 3.7 se representan los resultados para cada una de las cadenas con origen en el genoma nuclear. En la mayoría de los casos se obtuvieron valores menores a 1, con la salvedad de la cadena E (COX 5A), cuyo resultado fue mayor a 1.7, lo que indica una alta conservación de los residuos en contacto de esta subunidad. Sin embargo, debido al pequeño tamaño de estas subunidades (entre 46 y 148 residuos) y al menor número de secuencias disponibles (entre 14 y 30 especies), los resultados tan solo pueden aportar una idea cualitativa de su tendencia, puesto que quedan lejos de ser estadísticamente significativos. Ante tal dificultad, se realizaron otros análisis no paramétricos.

Para cada subunidad, se etiquetó cada residuo como “variante” o “invariante”, según su conservación en su correspondiente alineamiento múltiple (tabla 3.6). Por otro lado, el mismo residuo también se clasificó como “Contacto” o “No Contacto Expuesto”. Así, se procesó la proporción de residuos variantes, expuestos y libres (p_0) frente a la de variantes en contacto (p_1). Las diferencias entre estos valores ($d = p_0 - p_1$) deberían mostrar una distribución simétrica con media en 0 de acuerdo con la siguiente hipótesis nula: la variabilidad de los residuos es independiente de su rol en la interacción con otras subunidades. Para ensayar esta hipótesis se llevó a cabo una prueba de rangos con signo de Wilcoxon, mediante la que se rechazó H_0 con $p < 0.007$. Así, se demuestra que la alta proporción de residuos variables pertenecientes al grupo “No Contacto Expuesto” no es debida al azar. En otras palabras, existe dependencia entre las características estructurales llevadas a estudio y la tasa evolutiva.

No obstante, aunque los resultados de la prueba de Wilcoxon nos permitieron concluir que los residuos en contacto están más conservados en general, no sirve para detectar aquellas subunidades que destaque por su comportamiento evolutivo. Para ello, se construyeron tablas de contingencia de 2 x 2 para cada subunidad, enfrentando los dos criterios de clasificación explicados anteriormente (variabilidad y rol en la interacción). Bajo la hipótesis nula de que ambos criterios son independientes se llevó a cabo una prueba exacta de Fisher. Sólo se pudo rechazar H_0 con $p < 0.055$ para las subunidades COX 6C, 7A1, 7B y 7C (tabla 3.6), las cuales presentaban los mayores valores de d , es decir, las subunidades que presentan mayores diferencias en la variabilidad de los residuos en contacto con respecto a los libres y expuestos.

3. Evolución de residuos implicados en contactos intermoleculares

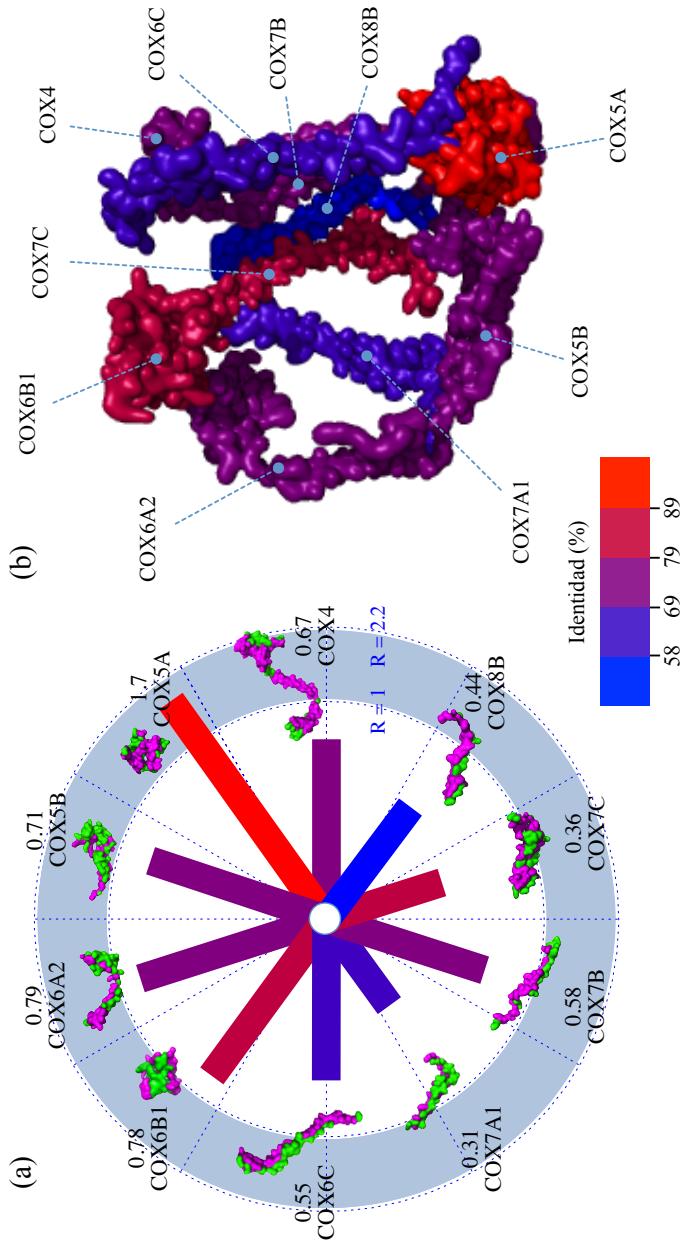


Figura 3.7: COX5A es la única subunidad codificada por nDNA que presenta una tasa de interacción $R_{\text{Contacto}, \text{No Contacto Expuesto}}$ superior a la unidad. En la gráfica (a), se representan los valores de las tasas de interacción y la estructura tridimensional de cada subunidad codificada por nDNA. Cada residuo de cada estructura ha sido representado de color magenta si está en contacto con otro residuo y en verde en caso contrario. En la imagen (b), se representa la estructura 3D de las 10 subunidades formando parte del complejo IV. El color de éstas es proporcional a porcentaje de identidad de secuencia en un alineamiento construido con aquellas especies para las que se dispone de todas las secuencias (*Ailuropoda melanoleuca*, *Bos taurus*, *Cavia porcellus*, *Equus caballus*, *Mus musculus*, *Rattus norvegicus*, *Sus scrofa*).

3.5. Resultados

Tabla 3.6: Relación entre la variabilidad de los residuos y su implicación en contactos con otras subunidades.

Subunidad	V-NCE	I-NCE	V-C	I-C	<i>p</i> -valor
COX 4	22	20	42	52	0.259
COX 5A	3	36	6	37	0.293
COX 5B	19	24	15	37	0.091
COX 6A2	16	20	21	27	0.561
COX 6B1	20	6	24	23	0.027*
COX 6C	19	8	9	22	0.002
COX 7A1	19	8	9	22	0.002*
COX 7B	11	6	9	23	0.015*
COX 7C	7	6	8	26	0.052*
COX 8B	10	4	17	12	0.320

Para cada subunidad de COX codificada por nDNA, se construyeron cuatro grupos de residuos atendiendo a la combinación de dos propiedades. Por un lado, se agruparon en “variantes” (V) o “invariantes” (I) dependiendo de si el residuo en cuestión varía o no en el alineamiento múltiple. Y por otro lado, se clasificó cada residuo atendiendo al grupo de residuos de superficie al que pertenece: “Contacto” (C) o “No Contacto Expuesto” (NCE). Para examinar la significatividad de la asociación entre estos dos criterios de agrupación, se llevó a cabo una prueba de Fisher. De esta forma, se calculó la probabilidad de observar por azar unos valores tan extremos o más extremos que los computados en las columnas 2-4, bajo las condiciones de la hipótesis nula (independencia entre los criterios de clasificación). Éstos valores se muestran en la columna 5 (*p*-valor).

* *p*-valor suficientemente bajo para rechazar la hipótesis nula y concluir que los residuos en contacto tienden a ser invariantes.

3.5.9 Correlación entre la edad de las proteínas y el valor de *d*

Con el objetivo de averiguar si las diferencias en la dinámica evolutiva entre los residuos implicados en contacto y los residuos libres están relacionadas con la edad de las proteínas, se llevó a cabo un sencillo análisis sobre estas variables. Las diferencias entre las proporciones de residuos variantes entre “Contacto” y “No Contacto Expuesto” de aquellas subunidades con valores significativamente mayores que cero, fueron representadas en una gráfica de dispersión frente a la edad de la correspondiente proteína (?). Se obtuvo una correlación negativa ($r = -0.946$, $p = 0.027$) estadísticamente significativa (figura 3.8). Estos resultados sugieren que las diferencias en la capacidad evolutiva entre residuos en contacto y residuos expuestos y libres tiende a ser máxima en proteínas más jóvenes.

3. Evolución de residuos implicados en contactos intermoleculares

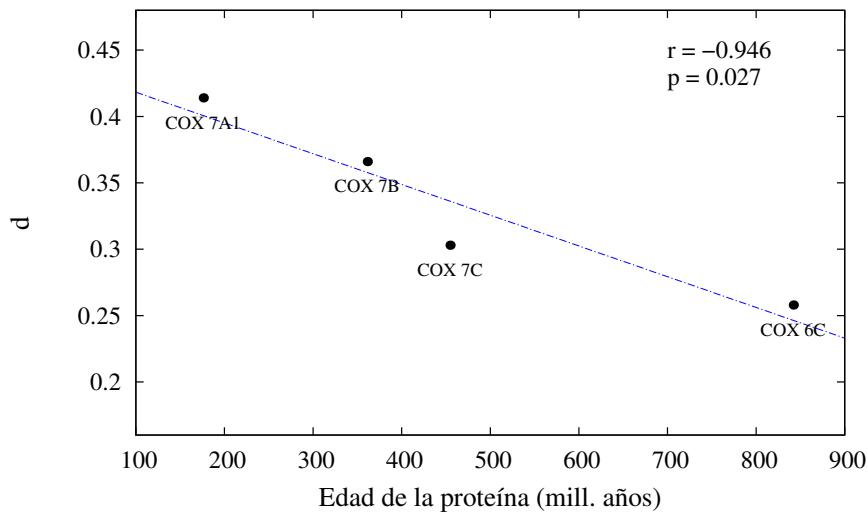


Figura 3.8: Las subunidades más jóvenes presentan mayores diferencias en capacidad evolutiva entre residuos en contacto y libres. Para cada subunidad se computó la diferencia en la proporción de residuos clasificados como “variantes” e ‘invariantes’. De las subunidades codificadas por nDNA, sólo COX 6C, 7A1, 7B y 7C mostraron valores de d significativamente mayores que cero. Al computar la edad de la proteína frente a los valores de d , se obtuvo una correlación negativa significativa.

3.6 Discusión

Es esperable que las interacciones físicas entre proteínas influyan de algún modo en la dinámica evolutiva de éstas. Tras el origen evolutivo de la unión entre dos proteínas, las regiones implicadas en el contacto entre dos o más cadenas de un complejo heteromérico, como citocromo c oxidasa, se han ido coadaptando, ya sea hacia la mejora de sus funciones fisiológicas (?) o simplemente manteniendo su eficacia biológica ante mutaciones levemente deletéreas que han sido fijadas por deriva genética (?). Caracterizar la coevolución de las distintas regiones en contacto del complejo IV de la cadena respiratoria, ha sido de especial interés, puesto que es un escenario en el que se encuentran interacciones entre subunidades codificadas por dos genomas que difieren en múltiples aspectos evolutivos. Como hemos detallado a lo largo de este capítulo, Schmidt y colaboradores (?), construyeron un modelo mediante el que estudiaron las tasas evolutivas de los aminoácidos implicados en la unión de las distintas subunidades del complejo citocromo c oxidasa, teniendo en cuenta sobretodo, la procedencia genómica de cada cadena. A partir de sus resultados, anunciaron la “hipótesis optimizadora”, que propone que los residuos codificados por mtDNA en contacto con otros codificados por nDNA evolucionan más deprisa que el resto de los aminoácidos codificados por el mismo genoma, favoreciendo la optimización de la unión (?).

A partir de las conclusiones alcanzadas en el capítulo 2 de esta tesis, mediante la

comparación de la evolución de los residuos expuestos frente a los enterrados de citocromo b y COX 1, surgieron dudas sobre la precisión del modelo de Schmidt y colaboradores (?). Dado que se encontraron diferencias significativas entre los residuos de superficie e interior (figura 2.7), consideramos que sería apropiado en este contexto, añadir un nuevo criterio de clasificación de residuos al modelo original. De este modo, las tasas evolutivas que se compararon, se estimaron en conjuntos de aminoácidos en contacto y en conjuntos de aminoácidos libres y expuestos, excluyendo a los enterrados para evitar sesgos en la muestra. Adicionalmente, en el modelo aplicado en este trabajo, se tuvieron en cuenta los residuos que se encuentran en contacto con cadenas del mismo origen genómico (figura 3.3).

3.6.1 Comparación de los dos modelos de cálculo de tasas de interacción

Los resultados obtenidos con el modelo de estudio propuesto en este trabajo, arroja múltiples evidencias en contra de la hipótesis optimizadora, y explican por qué los autores del modelo original (?) fallaron en la interpretación de sus resultados. Primero, la única subunidad con origen en el genoma mitocondrial, cuya tasa de interacción superó la unidad fue COX 1. El resto de las cadenas codificadas por mtDNA presentaron valores significativamente menores a 1 (tabla 3.4). Dado que en el modelo original no se tuvieron en cuenta las distintas subunidades de forma individual, sino que se hizo uso de un alineamiento producto de la concatenación de todas las subunidades del complejo, el excepcional comportamiento de COX 1 pasó desapercibido y produjo un sesgo en los resultados, el cual llevó una interpretación generalizada e incorrecta. Segundo, la exclusión de los residuos enterrados, así como aquellos que están implicados en contactos intragenómicos, redujo notablemente los valores de las tasas de interacción entre aminoácidos en contacto y los libres y expuestos para todos los casos (tabla 3.4). Esto indica que en el trabajo de Schmidt et al. (?) se subestimaron los valores de Σd_N del conjunto de residuos que ocupaba el denominador en el cálculo de la tasa de interacción (ecuación 3.2) entre residuos en contacto y libres, puesto que no se tuvo en cuenta la elevada conservación de los residuos del interior. Por último, el valor de Σd_N de los residuos expuestos y libres de COX 1 es inusualmente bajo, tanto si se compara con los de COX 2 y COX 3 (figura 3.4), como si se compara con el resto de residuos de la misma cadena (figura 3.5). Esto indica que es más probable que la causa de la alta tasa de interacción de COX 1 sea debida a la elevada conservación de los residuos expuestos y libres, más que a una alta tasa de cambio de los residuos en contacto. Así pues, estos resultados contrastan firmemente con la idea de que existe un proceso evolutivo de optimización de contactos entre las subunidades codificadas por distintos genomas en el complejo IV de la cadena respiratoria.

3. Evolución de residuos implicados en contactos intermoleculares

3.6.2 ¿Por qué los residuos expuestos y libres de COX 1 están tan conservados?

La contribución a la estabilidad y/o a la funcionalidad fueron las primeras variables que se consideraron como posibles causas que constriñen la evolución de los aminoácidos expuestos de COX 1. Por un lado, aunque se obtuvo una correlación negativa entre los cambios en la estabilidad y Σd_N , este conjunto difería en gran medida de los resultados esperados por la regresión lineal (figura 3.6), por lo que el mantenimiento de la estructura por parte de este conjunto no es una causa directa de su baja tasa de cambio. Con respecto a la implicación en la función de la proteína por parte de este grupo de aminoácidos, es cierto que se encuentra una mayor proporción de residuos implicados en la función del complejo (figura 3.6). Sin embargo, no creemos que esta interpretación sea concluyente, puesto que no hay certeza de que este conjunto de residuos, descrito en la literatura, esté completo. No obstante, hay indicios de que estabilidad y funcionalidad influyen en la conservación de los residuos expuestos y libres de COX 1. Así, es muy probable que la explicación real se encuentre en una combinación de ambas variables.

Otro asunto sobre el que se ha reflexionado, concerniente a los residuos expuestos y libres de COX 1, así como los conjuntos análogos en otras subunidades del complejo IV, es la formación de los llamados supercomplejos. Durante la pasada década, se han aportado evidencias experimentales de la existencia de una organización de la cadena respiratoria en un orden mayor de complejidad, conocida como *respirosoma* (?; ?; ?). En la misma línea, dos estudios recientes han publicado la estructura 3D del supercomplejo I₁III₂IV₁, resuelta por crio-microscopía electrónica con una resolución de 19-22 Å (?; ?). Estos modelos revelan algunos sitios donde complejos vecinos se encuentran suficientemente cercanos como para formar puentes salinos o enlaces de hidrógeno. Tres de estos sitios se encuentran entre los complejos III y IV, sin embargo, ningún residuo de COX 1 estaría implicado en estas uniones. Por otra parte, existen otras formas de supercomplejos tales como I₁III₂IV₂, I₁III₂IV₃ y I₁III₂IV₄, descritas en mitocondrias de corazón bovino (?). Dado que no se dispone de la estructura tridimensional de estas formas de ensamblaje, la contribución de los residuos de COX 1 en el respirosoma no se puede resolver completamente, pero es importante tener en cuenta, que puede influir en la dinámica evolutiva de ciertas regiones del complejo.

3.6.3 Comportamiento diferencial entre interacciones jóvenes y antiguas

De todos los residuos expuestos, aquellos implicados en la interacción entre distintas cadenas codificadas por mtDNA son los más conservados. Además, la comparación entre los dos grupos de residuos en contacto, intra- e intergenómicos (tabla 3.4), sugiere que aquellos residuos codificados por el genoma mitocondrial, en contacto con residuos del mismo origen genómico, están sujetos a mayores constricciones que aquellos que están implicados en interacciones con subunidades codificadas por nDNA. Hasta donde sabe-

mos, este es el primer estudio cuantitativo que da pie a tal conclusión, la cual se discute a continuación. En mamíferos, las tres subunidades codificadas por mtDNA forman el centro catalítico del complejo, mientras que las 10 subunidades codificadas por nDNA actúan como un caparazón alrededor del núcleo (?). En procariotas, el complejo citocromo c oxidasa está reducido al núcleo catalítico, el cual parece tener un origen ancestral (?). Por el contrario, las células eucariotas poseen genes nucleares que codifican para subunidades adicionales, cuyo número generalmente aumenta con la complejidad del organismo. Aunque ni el origen ni la función específica de estas cadenas no catalíticas está del todo revelado, parece ser que son significativamente más jóvenes que las proteínas que forman el núcleo (?; ?; ?). Por otra parte, se sabe que las proteínas jóvenes suelen estar sometidas a una selección purificadora más débil y evolucionan más rápido que las proteínas más antiguas (?; ?). Así, en este trabajo se propone, que si esta hipótesis es cierta para proteínas, también puede serlo para las interacciones. En otras palabras, es razonable asumir que para una proteína con una antigüedad determinada, aquellos residuos implicados en interacciones jóvenes evolucionan más rápido que aquellos implicados en interacciones antiguas. Esta hipótesis puede explicar las diferencias en las tasas evolutivas de los residuos en contacto con subunidades codificadas por nDNA y mtDNA (tabla 3.4). Aunque está fuera de los objetivos principales de este trabajo, sería interesante para el futuro, investigar si la relación entre el grado de selección y la edad de las interacciones, descrita aquí para las subunidades codificadas por el genoma mitocondrial del complejo IV, se reproduce también en otras proteínas.

Por otra parte, los cálculos de $\Delta\Delta G$ (tabla 3.5), indican que las diferencias evolutivas entre los contactos intra- e intergenómicos no puede ser explicada por la contribución de las uniones a la estabilidad termodinámica del complejo. Es posible que existan causas funcionales que estén detrás del alto grado de conservación de los residuos codificados por mtDNA en contacto con cadenas codificadas por el mismo genoma.

3.6.4 Comportamiento evolutivo de las subunidades codificadas por nDNA

Estudios anteriores han caracterizado la evolución de las subunidades del complejo IV codificadas por el genoma nuclear (?; ?; ?; ?). Sin embargo muchas de las conclusiones alcanzadas en estos trabajos están, aparentemente, en conflicto. Por ejemplo, Schmidt y colaboradores sostienen que los residuos codificados por nDNA, en contacto con otras subunidades están sujetos a una fuerte selección purificadora (?), mientras que Osada y Akashi llegaron a la conclusión de que hay indicios de selección positiva para estos residuos (?). Posiblemente, este conflicto pueda ser atribuible a diferencias en la metodología. De hecho, mientras los primeros autores usaron conjuntos de datos formados por residuos en contacto y libres, sin distinguir entre las distintas subunidades del complejo, los segundos usaron una técnica de detección de posiciones sometidas a selección positiva descrita previamente en la literatura (?). En nuestro trabajo, hemos abordado esta cuestión con un modelo que podría considerarse a medio camino entre los descritos

3. Evolución de residuos implicados en contactos intermoleculares

previamente, ya que consiste en analizar cada cadena codificada por el genoma nuclear de forma individual, aplicando el mismo modelo que se ha usado para las subunidades codificadas por mtDNA.

Los resultados representados en la figura 3.7 son, en líneas generales, coherentes con los obtenidos por Schmidt et al. (?), puesto que los residuos en contacto con otras cadenas evolucionan más lentamente que los libres y expuestos. Es cierto que, en este modelo, nos es potencialmente posible detectar subunidades que exhiben un comportamiento evolutivo particular con respecto al resto. Sin embargo, la escasez de datos de secuencias disponibles para estos genes, limitan la potencia estadística de nuestros resultados. Tan solo con respecto a las subunidades COX 6C, COX 7A1, COX 7B y COX 7C puede afirmarse de un modo significativo que existe un mayor número de residuos invariantes en las regiones en contacto con otras cadenas (tabla 3.6). Además, estas diferencias son más acusadas en proteínas más jóvenes (figura 3.8).

Una teoría propuesta recientemente, explica que la deriva genética de mamíferos, unida a bajos tamaños poblacionales, puede resultar en leves deficiencias estructurales que hacen a las proteínas más vulnerables a la desestabilización por el solvente, debido a la exposición de enlaces de hidrógeno. La formación de nuevas uniones proteína-proteína puede restaurar la estabilidad original (?). De acuerdo con esta hipótesis, los complejos heteroméricos presentes en mamíferos pudieron desarrollarse, no como respuesta adaptativa, sino como cambios compensatorios a favor de la estabilidad de la estructura y el mantenimiento la función ante cambios nocivos no deletéreos. Una vez que las nuevas proteínas se encuentran formando parte del complejo, puede darse un aumento de la selección purificadora sobre aquellos residuos que se encuentran en contacto, a diferencia de los residuos libres. Posteriormente, las subunidades recién añadidas pueden ganar funciones que serán de nuevo sometidas a presión selectiva, suavizando entonces las diferencias entre residuos en contacto y libres.

En un esfuerzo de presentar gráficamente esta propuesta, se ha representado en la figura 3.9 nuestra reflexión sobre la historia evolutiva de una proteína que forma parte de un complejo. Suponemos una proteína que no forma parte del complejo y que no tiene ninguna función. Tras el evento que supone el origen de la unión de dicha proteína al complejo proteico, y asumiendo que supone una mejora sobre la eficacia biológica del organismo, un grupo de residuos, implicados en dicho contacto, se verán sometidos a una selección purificadora creciente y, por tanto, la tasa de interacción disminuirá reflejando las diferencias en la dinámica evolutiva de la región de contacto y el resto de la superficie. Dado que la tasa de cambio de los residuos libres y expuestos es suficiente alta como para admitir como probable la adquisición de nuevas funciones, en el caso de que ocurra dicho evento, y suponga una mejora adaptativa, esta región se verá, del mismo modo que los residuos en contacto, sometida a selección purificadora, diluyendo las diferencias en la evolución de las distintas regiones de la proteína. El valor de R , dependerá entonces de la importancia del contacto interproteico y la función.

Se desconoce que las subunidades del complejo IV, codificadas por nDNA, desem-

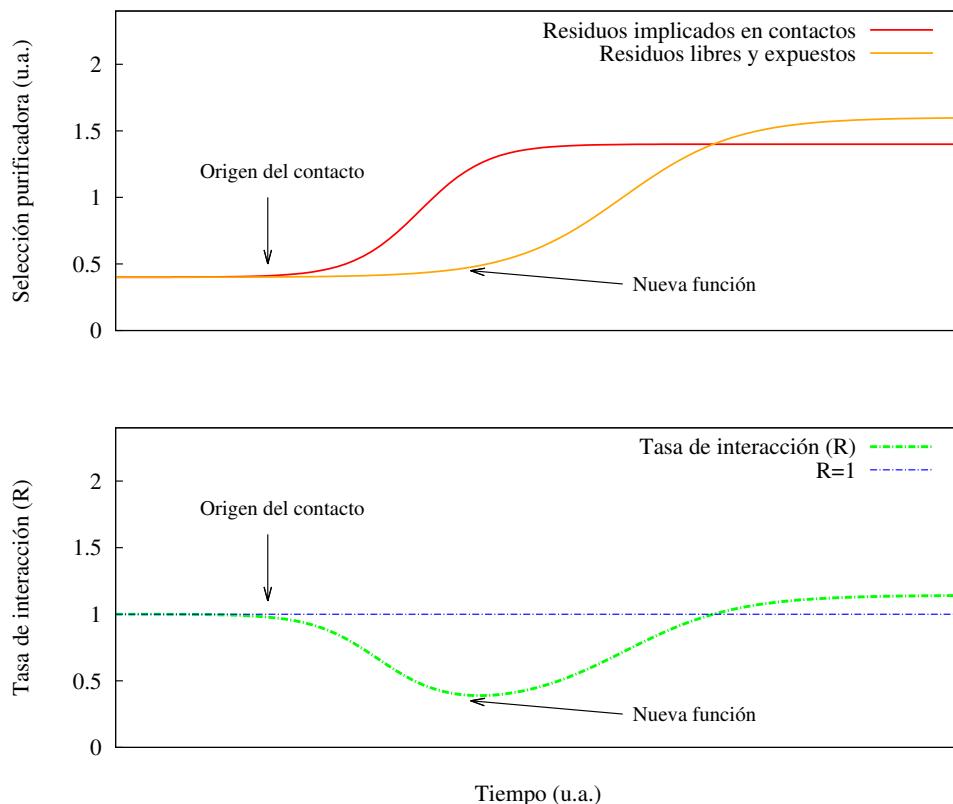


Figura 3.9: Relación de la tasa de interacción y la historia evolutiva de una subunidad codificada por nDNA. Al principio, una proteína que aún no forma parte del complejo, experimenta la misma magnitud de selección purificadora en toda su superficie, dando lugar a un valor de R cercano a la unidad. Una vez que se establece el contacto, los residuos implicados se ven sometidos a una mayor selección negativa, por lo que el valor de R disminuye. En el caso de que el resto de los aminoácidos de la superficie adquieran una nueva función que suponga una mejora adaptativa, ésta región tendería a conservarse, por lo que el valor de R aumentaría de nuevo pudiendo incluso superar la unidad.

3. Evolución de residuos implicados en contactos intermoleculares

peñen una función distinta a la propuesta por Fernandez y Lynch (?), que sugiere que el reclutamiento de nuevas subunidades por parte de un complejo proteico se debe, en primera instancia, a una mejora en la estabilidad del mismo. No obstante, a la subunidad COX 5A se le ha atribuido una función como receptor de 3,5-diyodotironina, un ligando capaz de eliminar la inhibición alostérica de la respiración por ATP (?). Según nuestra propuesta, COX 5A habría experimentado ambos eventos marcados con flecha, y se encontraría en la fase más tardía representada en la figura 3.9. Por el contrario, el resto de las subunidades codificadas por nDNA, cuyo valor de R es inferior a la unidad (figura 3.7), estarían en el intervalo de tiempo entre el origen del contacto y la posible adquisición de una nueva función.

3.7 Conclusiones

1. No existen evidencias de un fenómeno evolutivo que optimice las interacciones intermoleculares entre subunidades del complejo citocromo c oxidasa codificadas por distintos genomas (mtDNA y nDNA).
2. La tasa de sustituciones no sinónimas por sitio no sinónimo (Σd_N) es mayor en el caso de los residuos contacto mt-*nu* con respecto a los residuos contacto mt-mt, es decir, la evolución de los primeros es más rápida que la de los segundos.
3. La estabilidad termodinámica no explica el comportamiento evolutivo diferencial entre interacciones mt-mt y mt-*nu*.
4. Los residuos expuestos de COX 1, que no están en contacto con otras subunidades, están extraordinariamente conservados.
5. La estabilidad termodinámica correlaciona negativamente con la tasa evolutiva de los distintos conjuntos de residuos codificados por mtDNA, pero no explica, al menos totalmente, la excepcional conservación de los residuos No Contacto Expuestos de COX 1.
6. Se encuentran más residuos funcionales en el grupo de residuos No Contacto Expuesto de COX1 que en el grupo de residuos contacto. No obstante, no se dispone de suficiente potencia estadística para explicar la dinámica evolutiva del grupo de residuos “No Contacto Expuesto” mediante su funcionalidad.
7. Los residuos en contacto con subunidades codificadas por nDNA tienden a estar más constreñidos que los residuos no contacto de la misma cadena, salvo en el caso de la subunidad COX 5A.

Apéndice A

Mecom (*Molecular Evolution of protein COMplexes*)

A.1 Introducción

Una de las virtudes de la biología computacional es la alta reproducibilidad. Esto significa que no importa cuántas veces se repita un mismo ensayo que, bajo los mismos parámetros, se obtendrán siempre los mismos resultados. Así pues, se puede considerar que muestras distintas, sometidas a un mismo tratamiento computacional, darán resultados comparables entre si, con un error experimental infinitésimo. No obstante, este hecho no es tan evidente desde un punto de vista práctico. En numerosas ocasiones, el investigador se enfrenta a cuestiones cuya solución se halla tras un desarrollo de software de una complejidad considerable, y es en este proceso donde el error experimental puede tomar protagonismo. El motivo es obvio, cuanto más complejo es un modelo computacional, mayor es la probabilidad de que existan errores en algún punto, pudiendo pasar desapercibidos e introduciendo artefactos en los resultados.

Siendo conscientes entonces, de la relación positiva entre complejidad y vulnerabilidad de un modelo computacional, no es en vano la inversión de un esfuerzo adicional para tratar de aumentar, cuanto más mejor, la robustez del software desarrollado. Para ello, es buena práctica incluir dicho software en un proceso de desarrollo colaborativo, en el cual, según afirma en su libro, *The Cathedral and the Bazaar* (?), el informático americano y defensor a ultranza del movimiento *Open Source*, Eric Raymond, casi todos los errores serán caracterizados y arreglados rápidamente, siempre que exista un numero elevado de co-desarrolladores y *beta-testers*¹ implicados.

De acuerdo con estas prácticas, el software que ha sido desarrollado para llevar a cabo los análisis explicados en el capítulo 3 (página 58), ha sido publicado en repositorios

¹Literalmente: «*Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix will be obvious to someone*».

A. Mecom (*Molecular Evolution of protein COMplexes*)

públicos (véase sección A.4) bajo una licencia GNU GPL (*General Public License*) que permite el uso, la difusión y la modificación parcial o total del código, con el ánimo de hacerlo llegar a cualquier persona interesada en los objetivos que persigue el programa. Además, se ha hecho uso de un popular sistema de control de versiones (**Git**) que facilita el mencionado desarrollo colaborativo (véase sección A.6).

Así pues, Mecom, como hemos llamado al programa desarrollado en esta tesis como acrónimo de *Molecular Evolution of protein COMplexes*, y disponible en <http://mecom.hval.es/>, es un software de código abierto y de libre distribución destinado al análisis estructural y evolutivo de complejos protéicos. Mecom ha sido usado en este trabajo de investigación y su proyecto futuro incluye mejoras en su versatilidad y robustez y la aplicación en trabajos distintos enmarcados en la evolución molecular computacional. En este apéndice se describen los detalles técnicos del programa en su versión actual (v1.15), así como las plataformas sobre las que se desarrolla y distribuye, con el fin de proveer de la información suficiente a todo aquel que desee colaborar en este proyecto de software libre.

A.2 Implementación

Tal y como se explica de forma sucinta en el capítulo 3 (página 58), Mecom implementa una serie de análisis consecutivos sobre la estructura de un complejo protético y su evolución. Para ello, incluye algoritmos destinados al manejo de modelos atómicos (archivos *.pdb) y alineamientos múltiples de secuencias.

Mecom es la integración en un solo programa de un conjunto de scripts escritos en Perl y de otros dos programas (DSSP y PAML). Está construido de forma modular, es decir, el programa se divide en módulos que realizan funciones aisladas. En la tabla A.1 se resume la información correspondiente a cada uno de los módulos implementados en Mecom.

Los módulos pertenecientes a Mecom integran numerosas dependencias de otros módulos programados por terceros, disponibles en el repositorio público destinado a Perl por excelencia, CPAN (*Comprehensive Perl Archive Network*, <http://www.cpan.org/>). En los siguientes apartados se explica el funcionamiento de cada módulo (tabla A.1), en orden de ejecución.

A.2.1 Módulo Mecom::Contact

El módulo **Mecom::Contact** es el que realiza los cálculos que más tiempo consumen, por lo que, posiblemente, sea el módulo más susceptible a ser mejorado con el objetivo de aumentar la eficiencia del programa en su conjunto. Éste toma como datos de entrada el modelo atómico del complejo y calcula todas las distancias de todos los pares de grupos atómicos posibles. Posteriormente, toma la distancia más corta entre cada par

A.2. Implementación

de residuos pertenecientes a subunidades distintas del mismo complejo, y devuelve una lista de pares de residuos junto con la distancia computada. Por último, cada par de residuos se identifica como en contacto o no dependiendo de si el valor de proximidad es menor o mayor a un umbral preestablecido (por defecto 4Å).

La mejora más importante que necesita este módulo es la optimización del cálculo de la matriz de distancias.

A.2.2 Módulo `Mecom::Surface`

Este módulo toma como dato de entrada el archivo `.*pdb` que contiene el modelo atómico del complejo, al igual que el módulo `Mecom::Contact`, y realiza una llamada al sistema que consiste en el lanzamiento del programa DSSP. Posteriormente, recoge los datos de salida de DSSP entre los que se encuentran los valores de las superficies, en angstroms, de cada residuo de cada cadena del complejo. Por último, calcula el porcentaje de superficie expuesta de cada residuo tomando como 100 % el valor en angstroms de la superficie del aminoácido incluido en un triplete GXG totalmente extendido ($\Phi= -120^\circ$, $\Psi=140^\circ$), siendo X el aminoácido en cuestión y G glicina. Posteriormente, clasifica los residuos como expuestos o enterrados según si sus correspondientes valores de %ASA superan o no un umbral determinado por el usuario, que por defecto está configurado en 5 %. Los valores de superficie máxima de cada aminoácido se tomaron de Lee y Richards (?), y se muestran en la tabla A.2.

Finalmente, el módulo `Mecom::Surface`, integra en un solo archivo de texto los datos procedentes de la fase de clasificación en residuos en contacto o libres y los de clasificación en expuestos o enterrados. Éste archivo puede ser usado, opcionalmente, como datos de entrada de Mecom, de esta forma, la ejecución del programa puede saltarse los pesados cálculos estructurales.

A.2.3 Módulo `Mecom::Subsets`

Los dos módulos explicados hasta ahora, `Mecom::Contact` y `Mecom::Surface`, son los únicos encargados de realizar análisis sobre el modelo estructural del complejo proteico. La información necesaria para clasificar los residuos, según los criterios de proximidad a otros residuos y exposición al solvente, queda registrada tras la ejecución de éstos. El módulo `Mecom::Subsets` es el encargado de generar los conjuntos de cada uno de los grupos de residuos posibles tras la combinación de los dos criterios de clasificación (contacto y exposición). Estos conjuntos son trasladados al siguiente módulo, dando paso a la fase del algoritmo encargada de la manipulación y análisis de secuencias alineadas.

A. Mecom (*Molecular Evolution of protein COMplexes*)

Tabla A.1: Módulos de Perl implementados en Mecom.

Módulo	Función	Archivo
Mecom::Contact	Calcula las distancias entre residuos y determina si están en contacto o no basado en un umbral preestablecido.	lib/Mecom/Contact.pm
Mecom::Surface	Implementa el programa DSPP para el cálculo de superficies (%ASA).	lib/Mecom/Surface.pm
Mecom::Subsets	Construye los conjuntos de posiciones en base a los resultados de los análisis de la estructura.	lib/Mecom/Subsets.pm
Mecom::Align::Subset	Divide un alineamiento en varios <i>subalineamientos</i> a partir de conjuntos de posiciones.	lib/Mecom/Align/Subset.pm
Mecom::EasyYang	Integra el programa PAML para los análisis evolutivos (Σd_N).	lib/Mecom/EasyYang.pm
Mecom::Statistics::RatioVariance	Calcula la varianza de un cociente y realiza una prueba Z para evaluar la significatividad de los resultados.	lib/Mecom/Align/Subset.pm
Mecom::Report	Construye un informe en HTML con los resultados de los análisis.	lib/Mecom/Report.pm

Tabla A.2: Superficie máxima de cada aminoácido (?).

Residuo	Superficie (Å)	Residuo	Superficie (Å)
A	113	L	180
R	241	K	211
N	158	M	204
D	151	F	218
C	140	P	143
Q	189	S	122
E	183	T	146
G	85	W	259
H	194	Y	229
I	182	V	160

A.2.4 Módulo `Mecom::Align::Subset`

A partir de un alineamiento múltiple de secuencias nucleotídicas, el módulo `Mecom::Align::Subset`, reagrupa los tripletes y genera varios alineamientos nuevos que contienen los nucleótidos que codifican para los residuos agrupados en cada uno de los subconjuntos obtenidos en la ejecución del módulo anterior. En otras palabras, este módulo es el que fragmenta el alineamiento en otros alineamientos que codifican para regiones determinadas de la proteína.

Hace uso de los módulos de `BioPerl` (véase sección A.3) encargados del manejo de alineamientos múltiples y entrada y salida de datos.

Éste, al igual que el módulo `Mecom::Contact`, es un buen punto para situar el foco en los procesos de mejora del software, pero por distintas razones. Actualmente, el código implementado en este módulo es eficiente, sin embargo, dada la naturaleza de los datos de entrada, es indispensable incrementar su robustez, debido a que los alineamientos múltiples pueden ser muy diversos atendiendo a varios aspectos como la disparidad en las longitudes de las secuencias, el número y la distribución de *gaps* y las ambigüedades en los nucleótidos. Otro motivo es referente a la siguiente fase de la ejecución de `Mecom`, que incluye la ejecución de `PAML`, el cual es muy astringente en cuanto a la integridad de los alineamientos que computa, es decir, que cualquier error en las secuencias, impedirá la ejecución y detendrá `Mecom`. Así, este módulo podría integrar, en próximas fases de su desarrollo, un algoritmo de validación y depuración de alineamientos para evitar la detención de la ejecución.

A. Mecom (*Molecular Evolution of protein COMplexes*)

A.2.5 Módulo Mecom::EasyYang

El módulo Mecom::EasyYang es lo que en programación se denomina un *wrapper*. Esto es, un fragmento de código que maneja la entrada y salida de otro programa instalado en el sistema. Éste, es el encargado de ejecutar PAML, concretamente el programa yn00, que realiza los cálculos de las tasas de sustituciones entre cada par de secuencias de un alineamiento de entrada. Posteriormente, suma todos los valores de d_N obtenidos y obtiene el valor de Σd_N (véase capítulo 3, ecuación 3.1) para dicho alineamiento. El módulo procede de igual forma para cada alineamiento y devuelve los valores de Σd_N de cada uno junto con sus respectivas varianzas.

A.2.6 Módulo Mecom::Statistics::RatioVariance

Una vez que se han obtenido los valores de Σd_N y varianzas de todos los alineamientos de todas las regiones de la proteína, acotadas según los criterios estructurales, se procede a calcular las tasas de interacción $R_{x,y}$ (véase capítulo 3, ecuación 3.2) y sus varianzas. Dada la ecuación A.1,

$$R_{x,y} = \frac{\bar{x}}{\bar{y}} \quad (\text{A.1})$$

para calcular las varianzas de $R_{x,y}$, se empleó la ecuación A.2,

$$\text{Var}(R_{x,y}) = \frac{1}{\bar{y}^2} [\sigma_x^2 + R_{x,y}^2 + \sigma_y^2 - 2R_{x,y}\rho_{x,y}\sigma_x\sigma_y] \quad (\text{A.2})$$

donde \bar{y} es la media de y , σ_x y σ_y son las desviaciones típicas de x e y , respectivamente, y $\rho_{x,y}$ es el coeficiente de correlación entre x e y .

Dado que, en el trabajo de investigación para el que se desarrolló este programa, las condiciones de hipótesis nula asumían que $R_{x,y}$ no es distinto de la unidad, se implementó en este mismo módulo una prueba Z como método de contraste que permitiera averiguar si las desviaciones de los valores de $R_{x,y}$ con respecto a 1 eran significativas.

A.2.7 Módulo Mecom::Report

El último módulo que se ejecuta es el encargado de escribir un reporte en HTML que contiene toda la información de la ejecución, tanto los resultados de salida, como los parámetros y archivos de entrada.

A.2.8 Módulos Mecom y Mecom::Config

Mecom contiene otros dos módulos que no han sido enumerados en la tabla A.1 debido a que no están incluidos en ninguna fase de la secuencia de análisis. El primero, Mecom

(de igual nombre que el programa), es el que integra el resto de los módulos, los ejecuta en el orden correcto, gestiona toda la memoria donde se almacena la información, y la transmite de un módulo a otro. El segundo, `Mecom::Config`, contiene información referente a la instalación del programa.

A.3 Dependencias

Además de los programas DSSP y PAML, y los módulos integrados en el programa, Mecom integra otros módulos de terceros como dependencias. Muchas de estas dependencias están integradas en el paquete BioPerl (<http://www.bioperl.org/>) que consiste en un conjunto de módulos de Perl destinados al análisis de datos biológicos, como secuencias, alineamientos, estructuras, etc. Otras, pertenecen a la familia de módulos denominada `Statistics`, que implementan diversas herramientas estadísticas como el cálculo del coeficiente de correlación o la prueba Z.

Todos los elementos que componen Mecom, descritos hasta ahora, constituyen el denominado árbol de dependencias (figura A.1). Éstas deben estar instaladas correctamente en el sistema para que el programa funcione. La primera que debe ser instalada es el intérprete de Perl, que es el motor que ejecuta Mecom. En la web dedicada al programa, se puede encontrar información sobre las versiones de Perl disponibles y otros requisitos según el sistema operativo (<http://mecom.hval.es/install#req>).

A.4 Distribución e Instalación

Una de las grandes ventajas del código abierto (*open source*) es la posibilidad de reutilizar código, es decir, incluir en el software en construcción, fragmentos de código que han sido desarrollados previamente para cumplir determinadas funciones, por ejemplo, entrada y salida de datos, manejo de ficheros, herramientas matemáticas, etc. Por este motivo, la programación modular, típica de Perl, es una buena práctica orientada a compartir el código, siendo los módulos las unidades independientes y transferibles. Sin embargo, de cara al usuario, la instalación de todo el árbol de dependencias puede ser tediosa, y se requieren herramientas que asistan la integración correcta de todos los elementos en el sistema operativo. Para solucionar este problema, en el caso de los programas desarrollados en Perl, entra en juego un conjunto de módulos denominado CPAN. Brevemente, CPAN es un intérprete de comandos (consola) que se conecta a través de una conexión a internet al repositorio del mismo nombre, descarga todo el árbol de dependencias del programa solicitado, y lo instala en el sistema. Adicionalmente, integra otras tareas como las de compilación, gestión de manuales, etc., en los casos que fuera necesario. Mecom ha sido distribuido a través del repositorio CPAN, por lo que éste es accesible desde la consola de CPAN.

A. Mecom (*Molecular Evolution of protein COMplexes*)

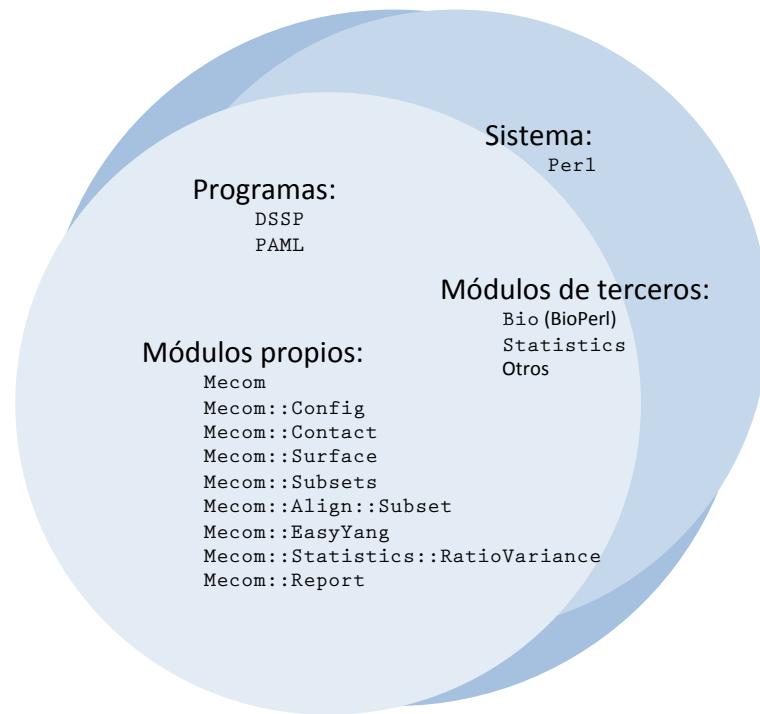


Figura A.1: Esquema conceptual de las dependencias de MECOM.

A.4.1 Instalación mediante CPAN

Instalar Mecom a través del sistema CPAN es el método recomendado por los autores debido a que, como se ha expuesto en el apartado anterior, se hace uso de un asistente responsable de descargar e instalar correctamente todo el árbol de dependencias. Para comenzar el proceso, es necesario tener instalado Perl y CPAN en el sistema. Con esto, la simple instrucción:

```
> perl -MCPAN -e 'force install Mecom'
```

descargará e instalará Mecom, incluyendo todas las dependencias, DSSP y PAML. Este último es también compilado por el mismo CPAN.

En la url <http://mecom.hval.es/install> se puede encontrar más información sobre la instalación.

A.5 Ejemplo de uso

Antes de comenzar a usar Mecom, es recomendable revisar el manual de usuario disponible en la web (<http://mecom.hval.es/manual>) o en el mismo intérprete de comandos. Si el programa está instalado correctamente, la instrucción

```
> man mecom
```

debe mostrar en pantalla un manual del programa. Del mismo modo, la instrucción

```
> man --help
```

muestra un resumen de los parámetros requeridos y opcionales.

El ejemplo más simple de uso de Mecom sería como sigue:

```
> mecom --pdb 2occ.pdb --chain M --alignment ChainM_alignment.fas
```

Esta instrucción está compuesta por el nombre del programa seguido de tres duplas. Cada una de éstas constituye un par “clave, valor”, donde la clave (precedida de dos guiones) es el nombre del parámetro que se desea declarar, y el valor es el contenido que se le desea asignar a dicho parámetro. Así, en el ejemplo, el primer parámetro, denominado `--pdb`, admite como valor la ruta al archivo que contiene el modelo atómico. El segundo, `--chain`, es el identificador de la subunidad para la que se desea realizar los cálculos y `--alignment` es la ruta al archivo que contiene los alineamientos.

Todos los resultados de la ejecución, se resumen en un archivo HTML que se escribe en la ubicación donde se ejecuta el programa. En él, se detallan los parámetros de entrada que se han utilizado, las rutas a los archivos donde se han almacenado los nuevos alineamientos, la información estructural y la evolutiva. Por último, muestra una tabla con los valores obtenidos de Σd_N y R de cada conjunto de residuos.

Más allá de este ejemplo, Mecom es altamente configurable, de modo que mediante el sistema “clave, valor” en las instrucciones de línea de comando, pueden modificarse todos los parámetros de la ejecución. Todos ellos se muestran en la tabla A.3. En el manual disponible en la web se ofrece una descripción más detallada de cada parámetro así como los valores por defecto de cada uno de ellos.

A.6 Desarrollo colaborativo

Como se ha señalado en la introducción (página 83), creemos que es necesario incluir este programa en un proyecto de desarrollo colaborativo. Para tal fin, hemos depositado todo el código fuente en una red de programadores denominada *Github* (<https://github.com/>). Esta plataforma, sumada al sistema de control de versiones *Git*, permite llevar a cabo este proceso de forma eficiente y cómoda. En la url <https://github.com/hvpareja/Mecom>, se encuentra el repositorio donde se ha depositado el código fuente, así como un foro interno en el que se discuten los distintos aspectos del desarrollo.

A. Mecom (*Molecular Evolution of protein COMplexes*)

Tabla A.3: Parámetros admitidos por Mecom en la línea de comandos.

Parámetro	Descripción
--help	Muestra un documento de ayuda.
--pdb†	Ruta al archivo *.pdb.
--contactfile§	Ruta al archivo de contactos ^a .
--alignment*	Alineamientos correspondientes a la secuencia nucleotídica de la cadena especificada.
--chain*	Cadena que se desea analizar.
--ocontact	Archivo donde se almacenarán los resultados de los análisis estructurales de la ejecución.
--exposureth	Umbral de exposición para determinar si el residuo está expuesto o enterrado.
--exposureerror	Margen de error del umbral de exposición.
--proximityth	Umbral de proximidad para determinar si un par de residuos está o no en contacto.
--informat	Formato de entrada del archivo de alineamientos.
--oformat	Formato de salida de los alineamientos parciales.
--gc	Código genético.
--report	Archivo donde se escribirá el reporte HTML.
--struct	Ejecuta sólo los análisis estructurales.

* Este parámetro es requerido para toda ejecución

† Este parámetro es requerido sólo en ausencia de --contactfile

§ Este parámetro es requerido sólo en ausencia de --pdb

^aArchivo que contiene la información estructural de una ejecución previa sobre la misma cadena.

Apéndice B

Modelos de barajado de secuencias

En el trabajo de investigación que se recoge en el capítulo 1 se desarrollaron distintos modelos computacionales con el objetivo común de obtener secuencias aleatorias de una frecuencia nucleotídica igual a la de las secuencias reales (véase página 15). En la sección de Resultados del capítulo 1 (página 17) sólo se incluyeron los resultados procedentes del barajado de secuencias mediante el modelo homogéneo, debido a que no se encontraron diferencias significativas entre los distintos modelos. En la figura B.1 se muestran los resultados que no fueron incluidos en el cuerpo de esta tesis.

B. Modelos de barajado de secuencias

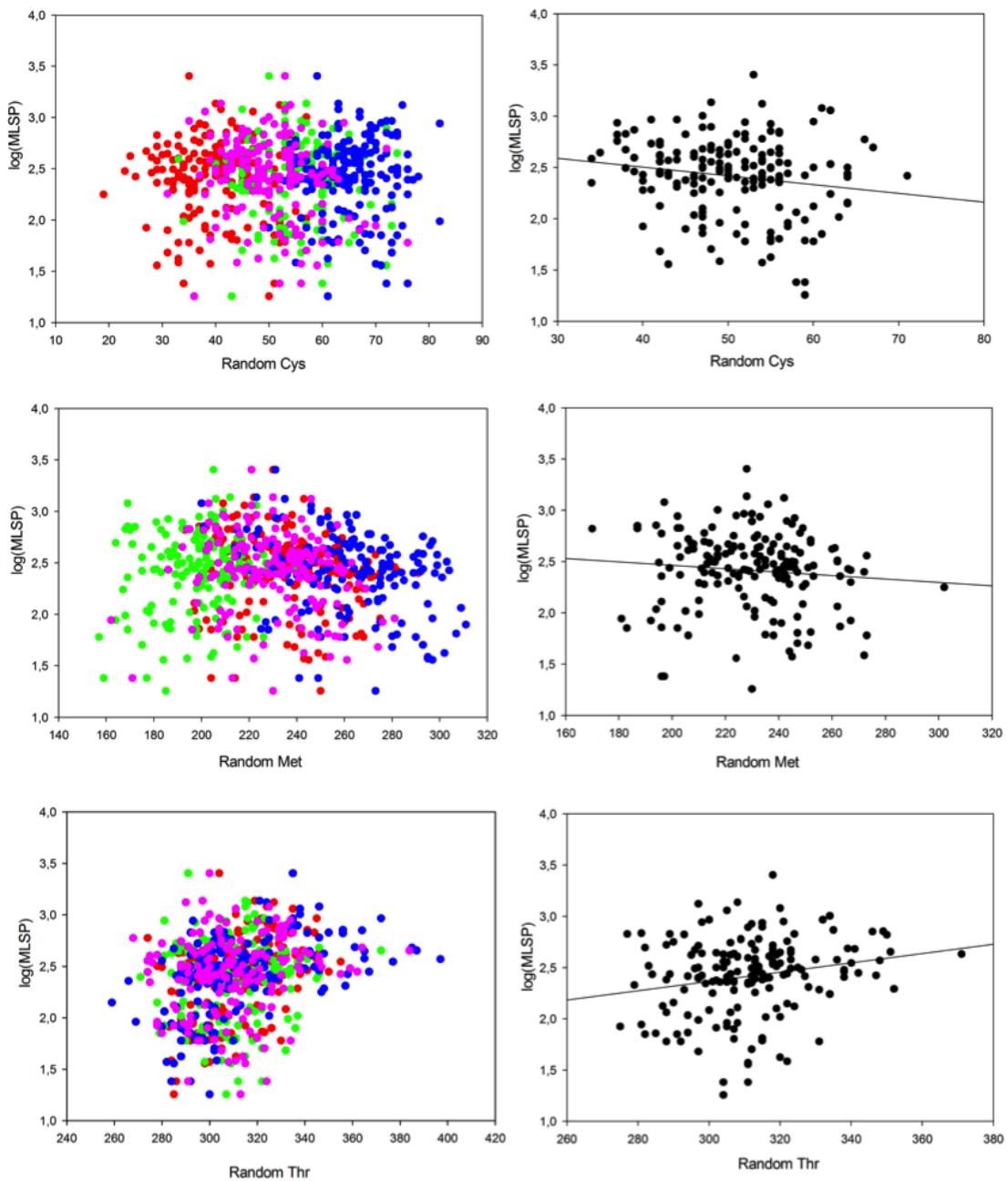


Figura B.1: Influencia de la composición nucleotídica sobre la abundancia de cisteína, metionina y treonina en proteínas codificadas por mtDNA. Para cada especie, se usó la secuencia, compuesta por los 12 genes concatenados de la hebra pesada, para generar otras secuencias aleatorias bajo los criterios de los distintos modelos, en los que cada posición de cada triplete es tratada como una categoría independiente. Los círculos rojos, verdes y azules se corresponden con los modelos 1, 2 y 3, respectivamente. Los círculos púrpuras representan los resultados obtenidos con el modelo 1-3. Por último, los círculos negros representan los resultados obtenidos con el modelo 1-2-3.

Apéndice C

Análisis de selección positiva en proteínas codificadas por nDNA

En el capítulo 1 se llevaron a cabo análisis sobre las proteínas codificadas por mtDNA para observar el efecto de la composición nucleotídica sobre el contenido de cisteína, metionina y treonina. De igual forma, en este apéndice se muestran los resultados de los mismos análisis aplicados sobre 61 proteínas codificadas por nDNA pertenecientes a una docena de especies de mamíferos (figura C.1). Se observa que, mientras $\Delta\text{Met} > 0$ en todas las especies analizadas, los valores de ΔCys y ΔThr fueron negativos en todos los casos. Estos resultados podrían sugerir la presencia de selección positiva sobre la abundancia de metionina también en el caso de las proteínas codificadas por nDNA. No obstante, dado el escaso número de especies computadas, no se pueden arrojar conclusiones sólidas.

C. Análisis de selección positiva en proteínas codificadas por nDNA

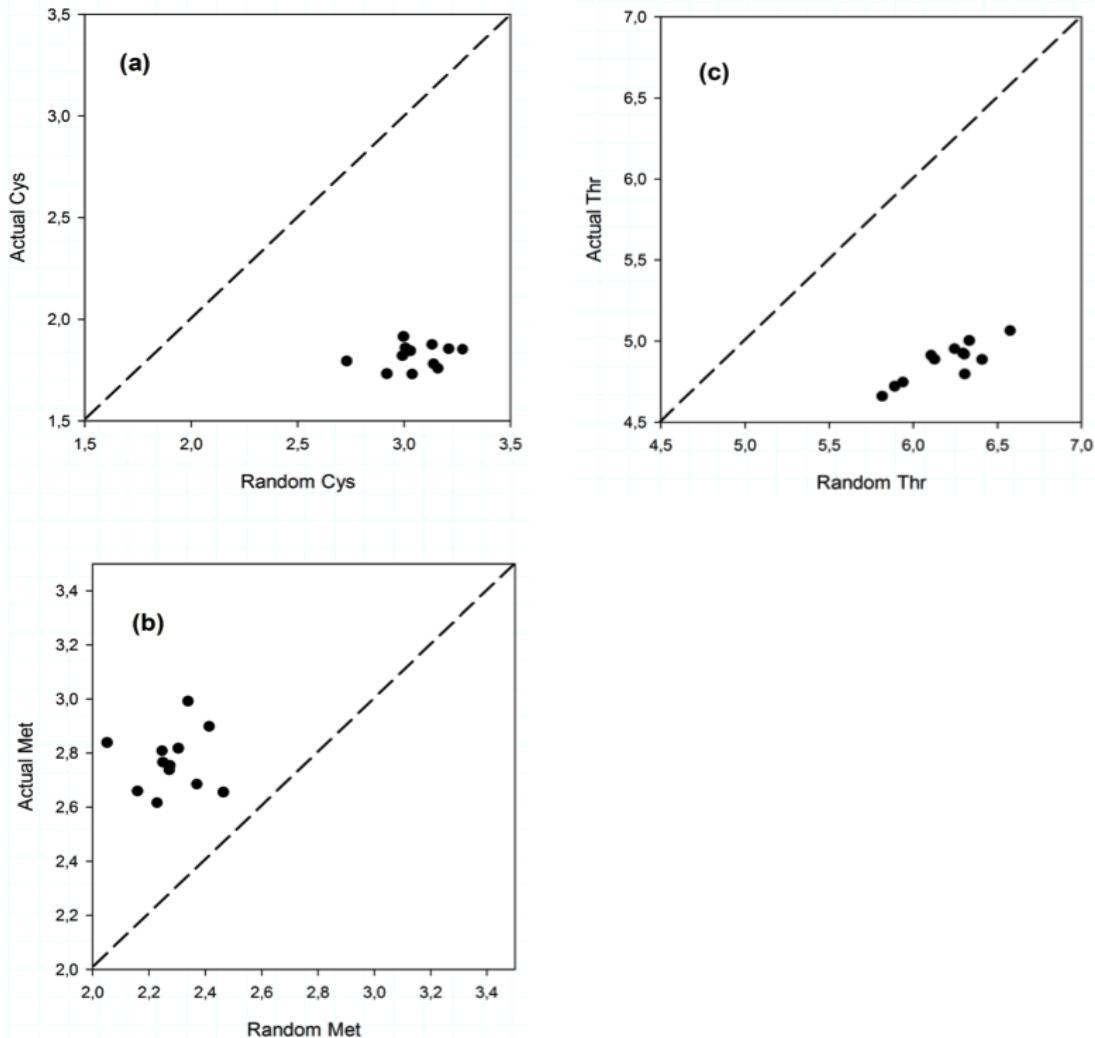


Figura C.1: Comparación entre la abundancia aminoacídica medida en las proteínas originales codificadas por nDNA y lo esperado por la única influencia de la composición nucleótídica. Para llevar a cabo este análisis se usaron 61 genes nucleares que codifican para proteínas pertenecientes a los complejos I, III y IV de la cadena transportadora de electrones de 12 mamíferos distintos (*Ailuropoda melanoleuca*, *Bos taurus*, *Canis familiaris*, *Equus caballus*, *Homo sapiens*, *Macaca mulatta*, *Monodelphis domestica*, *Mus musculus*, *Pongo abelii*, *Pan troglodytes*, *Rattus norvegicus*, *Sus scrofa*). Para cada especie, se computó el número de cisteínas (a), metioninas (b) y treoninas (c) presentes en las proteínas originales y se representó gráficamente frente al número de veces que el correspondiente aminoácido aparece codificado en la secuencia aleatoria, la cual presenta la misma composición nucleótídica que la original.

Apéndice D

Cálculo de las tasas de sustituciones de mtDNA

Tabla D.1: Pares de especies cercanas empleadas para el análisis de tasas de sustituciones en el genoma mitocondrial. En el capítulo 2 (página 36), se describe el proceso llevado a cabo para determinar los valores de d_N y d_S , a partir de las secuencias pertenecientes a las especies mostradas en esta tabla.

no.	Especie 1	Especie 2	Orden
1	<i>Chrysochloris asiatica</i>	<i>Eremitalpa granti</i>	Afrosoricida
2	<i>Ammotragus lervia</i>	<i>Capra hircus</i>	Artiodactyla
3	<i>Bos indicus</i>	<i>Bos taurus</i>	Artiodactyla
4	<i>Camelus bactrianus</i>	<i>Camelus dromedarius</i>	Artiodactyla
5	<i>Cervus elaphus</i>	<i>Cervus unicolor</i>	Artiodactyla
6	<i>Muntiacus muntjak</i>	<i>Muntiacus reevesi</i>	Artiodactyla
7	<i>Phacochoerus africanus</i>	<i>Sus scrofa</i>	Artiodactyla
8	<i>Arctocephalus forsteri</i>	<i>Phocarctos hookeri</i>	Carnivora
9	<i>Canis latrans</i>	<i>Canis lupus familiaris</i>	Carnivora
10	<i>Enhydra lutris</i>	<i>Lutra lutra</i>	Carnivora
11	<i>Eumetopias jubatus</i>	<i>Zalophus californianus</i>	Carnivora
12	<i>Halichoerus grypus</i>	<i>Phoca sibirica</i>	Carnivora
13	<i>Leptonychotes weddellii</i>	<i>Lobodon carcinophaga</i>	Carnivora
14	<i>Martes melampus</i>	<i>Martes zibellina</i>	Carnivora
15	<i>Panthera pardus</i>	<i>Panthera tigris</i>	Carnivora
16	<i>Phoca fasciata</i>	<i>Phoca groenlandica</i>	Carnivora
17	<i>Phoca largha</i>	<i>Phoca vitulina</i>	Carnivora
18	<i>Ursus arctos</i>	<i>Ursus maritimus</i>	Carnivora
19	<i>Balaenoptera acutorostrata</i>	<i>Balaenoptera bonaerensis</i>	Cetacea
20	<i>Balaenoptera brydeei</i>	<i>Balaenoptera edeni</i>	Cetacea
21	<i>Balaenoptera physalus</i>	<i>Megaptera novaeangliae</i>	Cetacea

D. Cálculo de las tasas de sustituciones de mtDNA

no.	Especie 1	Especie 2	Orden
22	<i>Berardius bairdii</i>	<i>Hyperoodon ampullatus</i>	Cetacea
23	<i>Eubalaena australis</i>	<i>Eubalaena japonica</i>	Cetacea
24	<i>Kogia breviceps</i>	<i>Physeter catodon</i>	Cetacea
25	<i>Monodon monoceros</i>	<i>Phocoena phocoena</i>	Cetacea
26	<i>Artibeus jamaicensis</i>	<i>Mystacina tuberculata</i>	Chiroptera
27	<i>Chalinolobus tuberculatus</i>	<i>Pipistrellus abramus</i>	Chiroptera
28	<i>Pteropus dasymallus</i>	<i>Pteropus scapulatus</i>	Chiroptera
29	<i>Rhinolophus monoceros</i>	<i>Rhinolophus pumilus</i>	Chiroptera
30	<i>Dasyurus hallucatus</i>	<i>Phascogale tapoatafa</i>	Dasyuromorphia
31	<i>Myrmecobius fasciatus</i>	<i>Sminthopsis douglasi</i>	Dasyuromorphia
32	<i>Metachirus nudicaudatus</i>	<i>Thylamys elegans</i>	Didelphimorphia
33	<i>Dactylopsila trivirgata</i>	<i>Petaurus breviceps</i>	Diprotontiae
34	<i>Phalanger interpositus</i>	<i>Trichosurus vulpecula</i>	Diprotontiae
35	<i>Phascolarctos cinereus</i>	<i>Vombatus ursinus</i>	Diprotontiae
36	<i>Pseudochirus peregrinus</i>	<i>Tarsipes rostratus</i>	Diprotontiae
37	<i>Erinaceus europaeus</i>	<i>Hemiechinus auritus</i>	Erinaceomorpha
38	<i>Dendrohyrax dorsalis</i>	<i>Procavia capensis</i>	Hyracoidea
39	<i>Ochotona collaris</i>	<i>Ochotona princeps</i>	Lagomorpha
40	<i>Ornithorhynchus anatinus</i>	<i>Tachyglossus aculeatus</i>	Monotremata
41	<i>Isoodon macrourus</i>	<i>Macrotis lagotis</i>	Paramelomorphia
42	<i>Ceratotherium simum</i>	<i>Rhinoceros unicornis</i>	Perissodactyla
43	<i>Bradypterus tridactylus</i>	<i>Choloepus didactylus</i>	Pilosa
44	<i>Colobus guereza</i>	<i>Procolobus badius</i>	Primates
45	<i>Eulemur fulvus fulvus</i>	<i>Eulemur fulvus mayottensis</i>	Primates
46	<i>Macaca sylvanus</i>	<i>Papio hamadryas</i>	Primates
47	<i>Nasalis larvatus</i>	<i>Pygathrix roxellana</i>	Primates
48	<i>Pongo abelii</i>	<i>Pongo pygmaeus</i>	Primates
49	<i>Tarsius bancanus</i>	<i>Tarsius syrichta</i>	Primates
50	<i>Elephas maximus</i>	<i>Mammuthus primigenius</i>	Proboscidea
51	<i>Microtus kikuchii</i>	<i>Microtus rossiaeemeridionalis</i>	Rodentia
52	<i>Rattus norvegicus</i>	<i>Rattus rattus</i>	Rodentia
53	<i>Sciurus vulgaris</i>	<i>Thryonomys swinderianus</i>	Rodentia
54	<i>Galemys pyrenaicus</i>	<i>Urotrichus talpoides</i>	Soricomorpha

Apéndice E

Publicaciones

E.1 Publicación del capítulo 1

Mutational Bias Plays an Important Role in Shaping Longevity-Related Amino Acid Content in Mammalian mtDNA-Encoded Proteins

Juan Carlos Aledo · Héctor Valverde ·
João Pedro de Magalhães

Received: 11 March 2012 / Accepted: 12 June 2012 / Published online: 30 June 2012
© Springer Science+Business Media, LLC 2012

Abstract During the course of evolution, amino acid shifts might have resulted in mitochondrial proteomes better endowed to resist oxidative stress. However, owing to the problem of distinguishing between functional constraints/adaptations in protein sequences and mutation-driven biases in the composition of these sequences, the adaptive value of such amino acid shifts remains under discussion. We have analyzed the coding sequences of mtDNA from 173 mammalian species, dissecting the effect of nucleotide composition on amino acid usages. We found remarkable cysteine avoidance in mtDNA-encoded proteins. However, no effect of longevity on cysteine content could be detected. On the other hand, nucleotide compositional shifts fully accounted for threonine usages. In spite of a strong effect of mutational bias on methionine abundances, our results suggest a role of selection in determining the composition of methionine. Whether this selective effect is linked or not to protection against oxidative stress is still a subject of debate.

Keywords Evolution · Longevity · Methionine · Cysteine · Threonine · Oxidative stress · Mitochondria

Electronic supplementary material The online version of this article (doi:[10.1007/s00239-012-9510-7](https://doi.org/10.1007/s00239-012-9510-7)) contains supplementary material, which is available to authorized users.

J. C. Aledo (✉) · H. Valverde
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain
e-mail: caledo@uma.es

J. P. de Magalhães
Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

Introduction

The free radical theory of aging asserts that buildup of macromolecular damage caused by reactive oxygen species (ROS) leads to the functional decline associated with aging in animals (Harman 2006). According to the current theory, mitochondria play a key role in aging because they are both a source of ROS (through leakage of the electron transport chain) and a major target for damage that could lead to a reduction in metabolic function (Balaban et al. 2005). Consistent with this theory, a number of mitochondrial adaptations to oxidative stress related to longevity have been described in the literature (for recent reviews, see Pamplona and Barja 2011; Pamplona 2011; Pamplona and Costantini 2011).

A first line of defense against ROS-related damage consists of ameliorating the rate of ROS production (Barja 1998). In this sense, comparative studies across-species have pointed to complex I of the electron transport chain (Pamplona et al. 2005; Lambert et al. 2007, 2010) and uncoupling proteins (Brand 2000; Speakman et al. 2004), as two potential targets of natural selection. Another strategy that animals seem to have adopted during the course of evolution consists of favoring macromolecular constituents less susceptible to oxidative modification. This point is well illustrated by the finding that the mitochondrial membrane peroxidizability index is inversely related to maximum life span in mammals (Pamplona et al. 1998). Along the same lines, different authors have suggested that concerted changes in mtDNA-encoded proteins leading to mitochondrial proteomes better endowed to resist oxidative stress may have evolved in response to selective pressures (Moosmann and Behl 2008; Kitazoe et al. 2008; Aledo et al. 2011).

Although all amino acid residues are potential targets of oxidative damage, the sulfur-containing residues cysteine

and methionine are particularly sensitive to oxidation (Berlett and Stadtman 1997). Interestingly, mitochondrialy encoded cysteine and methionine have both been correlated negatively with longevity (Moosmann and Behl 2008; Aledo et al. 2011). Even though both cysteine and methionine are negatively related to longevity, it may be that cysteine is a pro-oxidant while methionine is an anti-oxidant, as we have previously argued (Aledo et al. 2011). In brief, it is the detrimental capacity of cysteine thiyl radicals and their potential to initiate irreversible protein cross-linking that might have caused a selection against cysteine in mtDNA-encoded proteins. In contrast, the major oxidation product of protein-bond methionine is methionine sulfoxide, which can be reduced back to methionine by methionine sulfoxide reductases at the low metabolic cost of one molecule of NADPH (Stadtman et al. 2005). In this way, one equivalent of ROS is destroyed for every equivalent of methionine residue repaired (Levine et al. 1996). As such, a methionine enrichment of mitochondrial proteins may represent an adaptive response to oxidative stress (Bender et al. 2008; Aledo et al. 2011). Therefore, those animals that exhibit higher rates of ROS generation (short-lived animals) might have been subjected to higher selective pressures to increase the methionine content of their mitochondrial proteins. In other words, if mitochondrial methionine residues serve as a ROS sink, then the proteins from animals subjected to high oxidative stress should accumulate methionine more effectively than their orthologous proteins from species exposed to lower oxidative stress. That is, the relationship between methionine usage and longevity is somehow reminiscent of the well documented negative relationship between endogenous antioxidant levels and longevity (reviewed in Pamplona and Costantini 2011).

On the other hand, a positive correlation between threonine abundance in transmembrane regions of mitochondrial proteins and longevity has been reported (Kitazoe et al. 2008; Aledo et al. 2011). Since threonine is thought to provide extra intrahelical hydrogen bonding, thereby reinforcing protein stability, these authors argue that increased threonine content could be beneficial to achieve longer lifespan (Kitazoe et al. 2011).

Although these correlations of longevity with cysteine and methionine (negative) and threonine (positive) were originally interpreted in terms of Darwinian selection, Jobson et al. (2010) have recently raised doubts about the adaptive character of the above-mentioned amino acid compositional shifts. These authors point out that the reported correlations between longevity and amino acid usages might be explained by neutral mutational shifts of nucleotide composition and codon biases, which tend to be prominent in mitochondrial genomes (Albu et al. 2008). Indeed, a plethora of studies have established that protein

evolution is affected by the nucleotide composition of the encoding genes (Sueoka 1961; D'Onofrio et al. 1991; Collins and Jukes 1993; Singer and Hickey 2000; Wang et al. 2004). Mitochondrial genomes are not an exception to this general rule (Foster et al. 1997; Nikolaou and Almirantis 2006; Min and Hickey 2007; Jia and Higgs 2008; Jobson et al. 2010). In fact, it has been documented that uncorrected nucleotide bias in mtDNA can mimic the effects of positive selection (Albu et al. 2008). These precedents emphasize the difficulties of distinguishing between functional constraints in protein sequences and mutation-driven biases in the composition of these same sequences. This means that extreme caution should be used when comparing results between taxa that differ in their nucleotide contents, at the same time that it underscores the need for methodologies allowing to discriminate between neutral and selective forces.

In the current study, we have addressed the potential contribution of natural selection to the observed link between amino acid compositional shifts in mitochondrial proteomes and longevity. To this end, we have developed a framework that accounts for the effects of nucleotide compositional biases.

Materials and Methods

Data Sampling

Sequences for mitochondrial genomes and proteome analysis were obtained from the National Center for Biotechnology Information (NCBI) genome database (www.ncbi.nlm.nih.gov). A collection of 173 mammalian species was assembled based on neutral selection criteria, completely unrelated to longevity or nucleotide/amino acid frequencies. Annotated complete mitochondrial sequences had to be available from NCBI genome database, and longevity information had to be given in a reliable source such as the AnAge database at <http://genomics.senescence.info/species/> (de Magalhães and Costa 2009).

Computations

The sense sequences (L-strand) corresponding to the 12 protein-coding mitochondrial genes, without stop codons, were concatenated into a single nucleotide sequence. This sequence was translated into amino acid sequence using the vertebrate mitochondrial code. Six different nucleotide random sequence models were generated using a customized Perl script. In the so-called homogeneous model, only two constraints were imposed: (i) the nucleotide frequencies in the random sequence should be the same as in the actual sequence, and (ii) no stop codons were allowed. When a

stop codon appeared during the randomization procedure, the nucleotides from that triplet were, in turn, randomly shuffled until they coded for an amino acid. Beside this homogeneous model, we also considered modeling approaches where codon positions were treated as separate categories. For instance, for the model 1-2-3 the nucleotides at the three positions were shuffled in such a way that the frequencies at the first, second, and third position in the random sequence should be the same as the frequencies at the first, second, and third codon position of the actual sequence, respectively. In the models 1, 2, and 3, only the first, the second or the third position was shuffled, respectively. Finally, in the model 1-3, both the first and third positions were shuffled in such a way that the nucleotide frequencies at each position in the random sequence were the same as in the actual sequence. The calculations of the nucleotide and amino acid abundances were carried out with simple Perl code. All the scripts are available from the authors on request.

Statistical Treatment

For each species (*i*), the occurrences of the amino acid under investigation (methionine, cysteine or threonine) in the actual (x_i) and random (y_i) sequences were computed, yielding a pair of measurement (x_i, y_i) for each amino acid. In the absence of functional constraints in the protein sequences, the null hypothesis states that x_i and y_i are equally likely to be larger than the other ($P[X < Y] = P[X > Y] = 0.5$). Pairs were omitted for which there was no difference. In this way, the number of valid pairs was $n = 171$ for the methionine analysis and $n = 167$ for the case of threonine. Then, the number of pairs W for which $x_i - y_i > 0$ was calculated. Under the null hypothesis conditions, W follows a binomial distribution, $W \sim B(n, 0.5)$. However, since n is large enough, the normal approximation to the binomial distribution can be used. For this purpose, the variable W was typified taking into consideration that the mean is given by $n/2$ and the variance is $n/4$ ($E[W] = np = n/2$; $\text{Var}[W] = np(1-p) = n/4$). The typified variable, let us call it Z , follows now a normal distribution, $Z \sim N(0, 1)$.

Probability calculations were assisted by Wolfram Mathematica 8.0. All other statistic analyses were done with SPSS 15.0.

Results and Discussion

Since it has been suggested that the link between longevity and amino acid frequencies in mtDNA-encoded proteins may be a consequence of nucleotides biases rather than

reflect an adaptive process, we started by addressing the relationship of nucleotide abundances to maximum life span in a set of 173 mammalian species (Fig. 1). While adenine and guanine abundances showed weak, if any, relationships with longevity (p values: 0.001 and 0.070, respectively), thymine and cytosine abundances exhibited a strong relationship with lifespan (p values: 7×10^{-15} and 10^{-12} , respectively). The clear pattern in these data is that increases in longevity are accompanied by increases in the frequencies of C, which are paralleled by decreases in the frequencies of T. A similar trend has previously been reported (Samuels 2005). In this study 76 mammalian species were analyzed in the context of a differential susceptibility of mtDNA to damage, which according to this author may be brought about by nucleotide composition and be related to lifespan. While we confirm and extent this initial observation using a much wider set of species, that was not the main focus of the current paper.

It seems obvious that these nucleotide biases, which may be due to differences in generation times since long-lived species also have longer generation times and thus mutation biases will be skewed according to generation times, may bring about biases in the amino acid composition of the encoded proteins. Therefore, we next asked if the observed correlation between nucleotide composition and longevity may account for the previously reported links between longevity and cysteine, methionine and threonine usages in mtDNA-encoded proteins. To this end, for each species, the sense sequences encompassing the 12 proteins encoded by the H-strand were used to generate random sequences with the same nucleotide frequencies as the actual sequence. After translating these random sequences, the number of times a given residue appeared was plotted against the longevity of the species being analyzed. The results of such analyses showed that the nucleotide bias might be a significant driving force in methionine and threonine composition of the encoded proteins (Fig. 2). At this point, it may be argued that using a homogeneous model where the overall mitochondrial sequence is randomized, any constraints depending on codon position may be overlooked. Therefore, we next considered modeling approaches where codon positions were treated as separate categories of sites (see “Materials and Methods” section). The results derived from such models were qualitatively the same as those obtained with the homogenous model (see Supplementary Fig. 1). This congruency between models was not unexpected since Jobson et al. (2010) noted that the covariance between nucleotide usages and longevity showed a remarkable similar pattern regardless of the codon position, which was interpreted by these authors as evidence that the overall mitochondrial compositional pattern is driven by mutation biases.

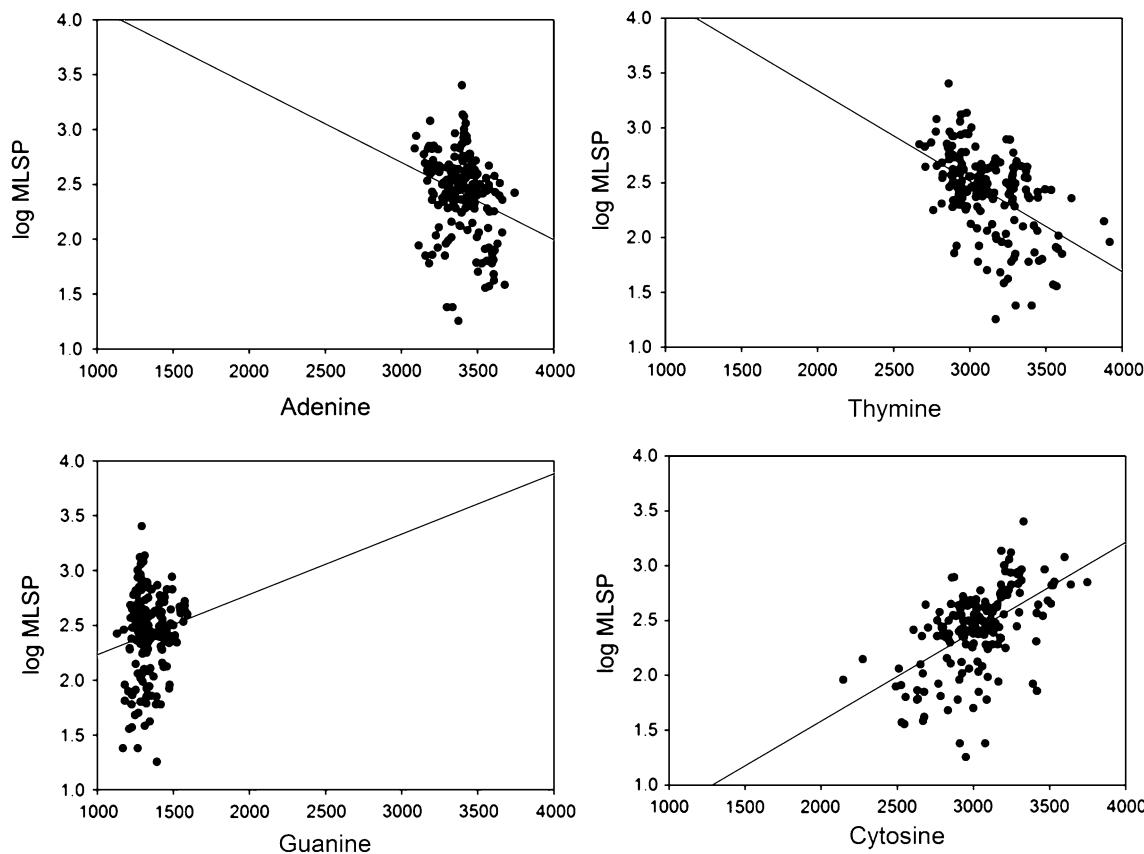


Fig. 1 Relationships between nucleotide abundance on the mtDNA L-strand and longevity. The 12 protein-coding genes on the L-strand from mtDNA were used to compute the base content. Then, the correlations between these absolute frequencies and log MLSP were addressed, using the data set formed by the 173 species analyzed in

the current study. Thymine and cytosine abundances exhibited a strong relationship with lifespan (p values: 7×10^{-15} and 10^{-12} , respectively), while adenine and guanine exhibited little, if any, relationship with longevity (p values: 0.001 and 0.070, respectively)

Whatever the underlying causes, nucleotide bias seems to play an important role in determining the methionine and threonine content. However, the question we wanted to answer was: Does nucleotide composition fully account for the observed amino acid usages? In other words, can we rule out selective forces contributing to shape the methionine and threonine composition of mtDNA-encoded proteins?

As a first approach to the above questions, we computed the number of times a given residue appeared in the 12 proteins encoded by the H-strand, plotted it against the number of times this residue appeared in a random sequence with abundances for the four bases equal to those observed for the species being considered (Fig. 3). As expected, there was a positive covariance between actual and random occurrences for the three amino acids. However, each amino acid behaved in a different way. As such, to evaluate the influence of nucleotide composition on each amino acid usage, two aspects need to be considered. On one hand, the slopes in these plots inform us about the inertia of amino acid usages with respect to mutational

bias. On the other hand, the distribution of the points with respect to the bisectrix can also be informative.

The nearly null slope exhibited by cysteine suggests the existence of strong constraints that buffer changes in the abundance of cysteinyl residues. Furthermore, the low variability of the actual cysteine usage variable, particularly true within long-lived animals (Fig. 3a), also suggests the action of a strong purifying selection. Even more, the fact that all the points lie below the bisectrix, points to a strong selection against cysteine. However, no effect of longevity on the departure from neutral mutation could be detected for this amino acid (Fig. 4a). In other words, we found that although cysteinyl residues were actively avoided, there were no differences between short- or long-lived species. This observation contrasts with previous results reporting a negative relationship between cysteine usage and longevity, across a wide range of animals covering mammals, birds, reptiles, amphibians, fishes, insects, crustaceans, and arachnids (Moosmann and Behl 2008). Nonetheless, when the correlation analyses were focused on the class Mammalia after correction for the effects of

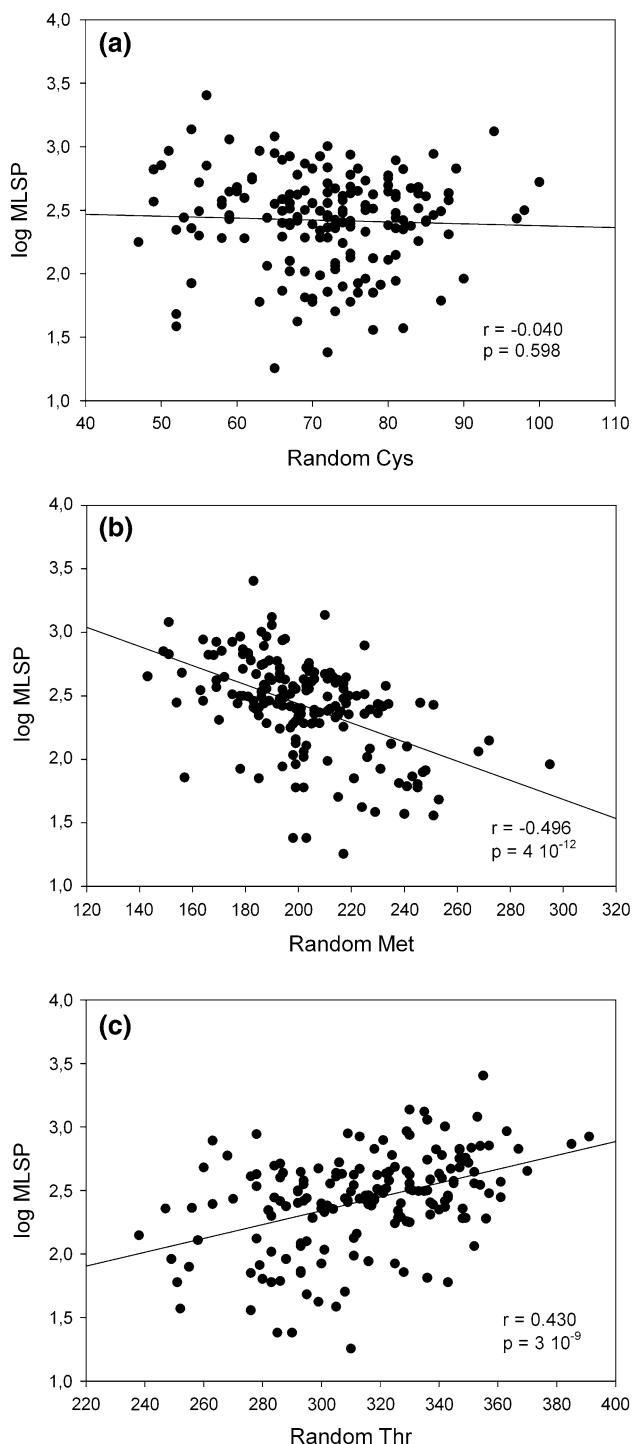


Fig. 2 Potential influence of the nucleotide composition on the cysteine, methionine, and threonine usages in mtDNA-encoded proteins. For each species, the sense sequence encompassing the 12 proteins encoded by the H-strand was used to generate random sequences with the same nucleotide frequencies than the actual sequence. Afterward, these random DNA sequences were translated using the vertebrate mitochondrial code and the number of times a cysteine (a), methionine (b) or threonine (c) residue appeared was plotted against the longevity of the species under analysis

phylogeny, no significant correlation between longevity and cysteine usage could be found (Aledo et al. 2011), which is line with the current observations (Fig. 4a). Thus, a note of caution should be sounded concerning a possible link between longevity and cysteine abundance in mtDNA-encoded proteins.

On the other hand, threonine frequency showed a great variability that seemed to be well accounted by the nucleotide composition, as indicated by a high slope (Fig. 3c). In addition, the points are equidistantly distributed around the bisectrix. These results, taken all together, suggest that threonine usages in mtDNA-encoded proteins are shaped by mutational bias and the corresponding nucleotide composition of the coding genes. These observations, while not ruling out the adaptive character of increased threonine usages in long-lived mammals (Kitazoe et al. 2008), throw doubts on it.

With regard to methionine usages, we observed a behavior that was intermediate between that described for cysteinyl residues and that noted for threonyl residues. That is, although the response to nucleotide composition seemed to be less constrained than that observed for cysteine, a certain degree of inertia became evident when compared to the response of threonine frequencies. Nevertheless, the fact to be emphasized is that most species carried a higher number of methionyl residues into their proteins with respect to that expected from the influence of nucleotide bias (Fig. 3b). This conclusion was quantitatively confirmed by performing further analyses.

For a given amino acid, the differences between the numbers of occurrences in the actual and random sequences were computed for each species. These differences (ΔAa) can be interpreted as a measurement of the departure from neutral mutational effects. For instance, the distribution of ΔThr was centered around zero, indicating that threonine frequencies are mainly shaped by random forces with little constraints. In contrast, ΔCys took negative values for all the analyzed species without exception, which suggests a selection against this amino acid. On the other hand, ΔMet showed a distribution centered around 30, with only a few species exhibiting negative values, indicating a certain degree of positive selection favoring methionine incorporation into mitochondrial proteins (Fig. 4a).

To substantiate this claim, we next carried out a sign test. For this purpose, we defined a random variable as the number of $\Delta Aa > 0$ occurrences. In this way, the null hypothesis states that in the absence of constraints, the typified random variable should follow a normal distribution, $Z \sim N(0, 1)$. While for threonine the null hypothesis was accepted (Z statistic = 0.0, p value = 0.5), it was clearly rejected for methionine (Z statistic = 9.2, p value = 2×10^{-20}). Since most mtDNA-encoded proteins use the AUG start codon, which encodes for methionine, ΔMet may well be inflated

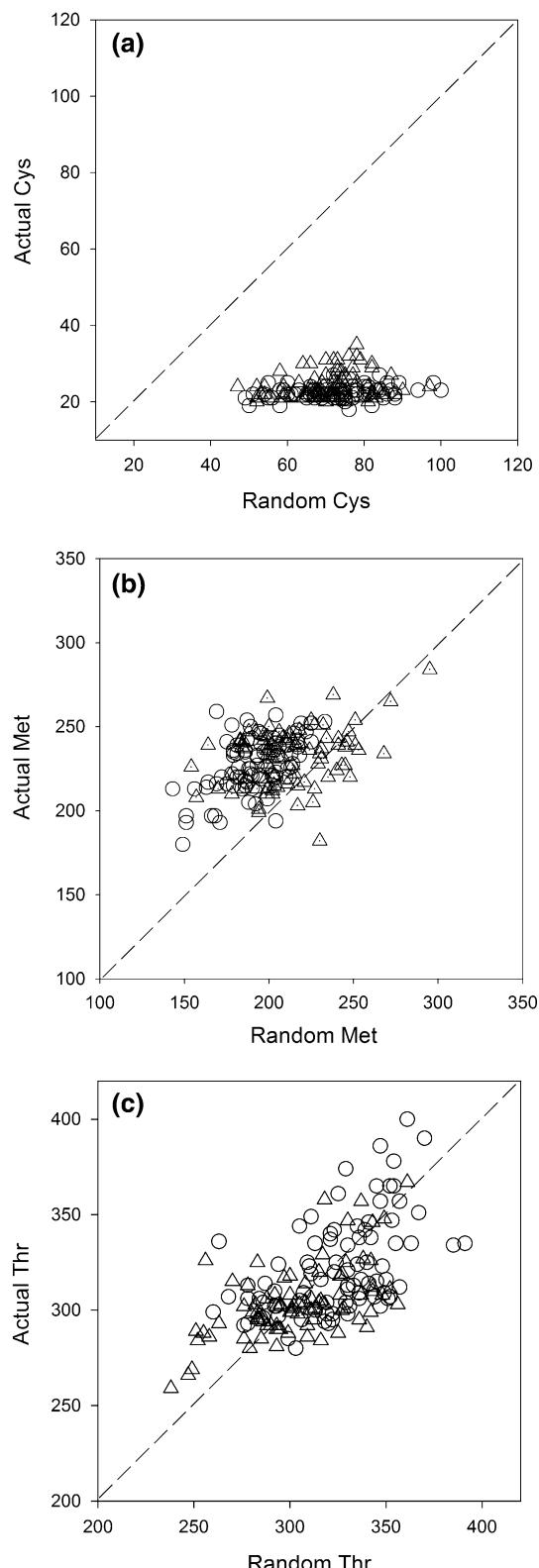


Fig. 3 Departures of the actual amino acid usages from random expectations. For each of the 173 mammalian species analysed in the current work, the number of cysteine (a), methionine (b) and threonine (c) residues present in the 12 concatenated proteins encoded in the same mtDNA-strand, were computed and plotted against the number of times that the corresponding amino acid appeared in a random sequence with abundances for the four bases equal to those observed for the species under consideration. We divided the whole sample into two groups, with short-lived mammals (*triangles*) having $\log \text{MLSP} < 2.46$ and long-lived mammals (*circles*) having $\log \text{MLSP} > 2.46$, where 2.46 (corresponding to 228.4 months) is the median of $\log \text{MLSP}$

translational effect. Nevertheless, once this translational effect was discounted, the null hypothesis that the typified random variable Z follows a normal distribution centered at zero, was still rejected ($Z = 6.6$, p value = 2×10^{-11}). These observations suggest that mtDNA-encoded proteins may incorporate more methionyl residues than expected from the neutral influence of nucleotide bias (Fig. 4b), which would be in line with the anti-oxidant role proposed for this amino acid (Levine et al. 1996; Bender et al. 2008; Aledo et al. 2011). Nevertheless, we acknowledge that further work and corroborating evidences will be required to fully support such a hypothesis.

A remarkable observation was that among the few species (21 species out of 173) that did not carry more methionine into their proteins than expected from random (those species that are distributed below the bisectrix from Fig. 3b), all, but one, belonged to the group of short-lived mammals (species with longevities below the median, $\log \text{MLSP} < 2.46$). Although we do not know the reasons behind this intriguing result, one is tempted to speculate that because the sense mtDNA-strand from long-lived mammals exhibit lower frequencies of T and higher frequencies of C (Fig. 1), which seems to favor the presence of the codons ACA and ACG (coding for threonine) at expenses of AUA and AUG (coding for methionine), long-lived animals are more in need of mechanisms, other than mutational bias, leading to increased methionine usages. To investigate this working hypothesis, we next addressed the correlation between ΔAA and longevity. As it can be observed in Fig. 5, there was a highly significant correlation between ΔMet and $\log \text{MLSP}$ ($r = 0.456$, p value = 3×10^{-10} , $n = 173$). In contrast, ΔCys and ΔThr did not show any relationship with longevity (p values = 0.3 and 0.5, respectively).

The lack of an association of ΔThr with longevity is congruent with the conclusion that threonine usage is mainly determined by the nucleotide composition of the considered mitochondrial genome. On the other hand, the lack of correlation between ΔCys and longevity could simply be due to an inverse ceiling effect. To this respect, it should be noted that mtDNA-encoded proteins exhibit a remarkable cysteine avoidance when compared to their

because of this translational constraint. Therefore, we repeated the analyses, this time excluding start codons. As it can be seen in Fig. 4a, a non-negligible fraction of the described methionine excess could be accounted for the

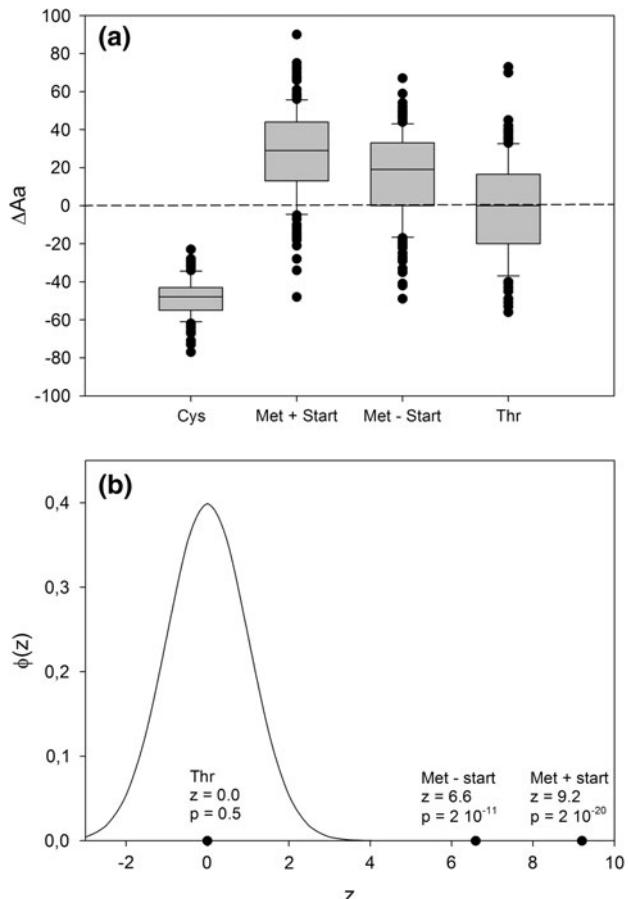


Fig. 4 mtDNA-encoded proteins incorporate more methionyl residues than expected from the influence of nucleotide bias. The differences between the numbers of occurrences of given residue in the actual and random sequences were computed for each species. The distribution of such a variable, ΔAa , is shown in **a**. While ΔThr is distributed around zero indicating the absence of selective constraints, ΔCys and ΔMet depart from random in opposite directions. Since most mtDNA-encoded proteins use the AUG start codon, which encodes for methionine, we carried out the analyses either including (Met + start) or excluding (Met - start) start codons. **b** The number of species for which methionine frequency is higher in the actual sequence with respect to its random sequence, can be considered as a random variable, which after typification yield the so-called Z statistic. In the absence of constraints, the Z statistic follows a normal distribution (see technical details in the “Material and Methods” section). The calculated value of the Z statistic for methionine was 9.2 when start codons were included in the analysis, or 6.6 when these start codons were excluded. Thus, the probability of the observed departure of ΔMet from random happening by chance is less than 2×10^{-11} . For comparative purposes, the Z statistic and p values for threonine were also calculated and indicated in the figure

nDNA-encoded counterparts (relative frequencies around 0.6 and 1.7 %, respectively). Furthermore, evolutionary pressure towards reduced global cysteine usage cannot essentially affect functionally indispensable residues, such as those that bind the numerous iron–sulfur clusters of the respiratory chain. Therefore, it may be that current mtDNA-encoded proteomes have reached a limit beyond

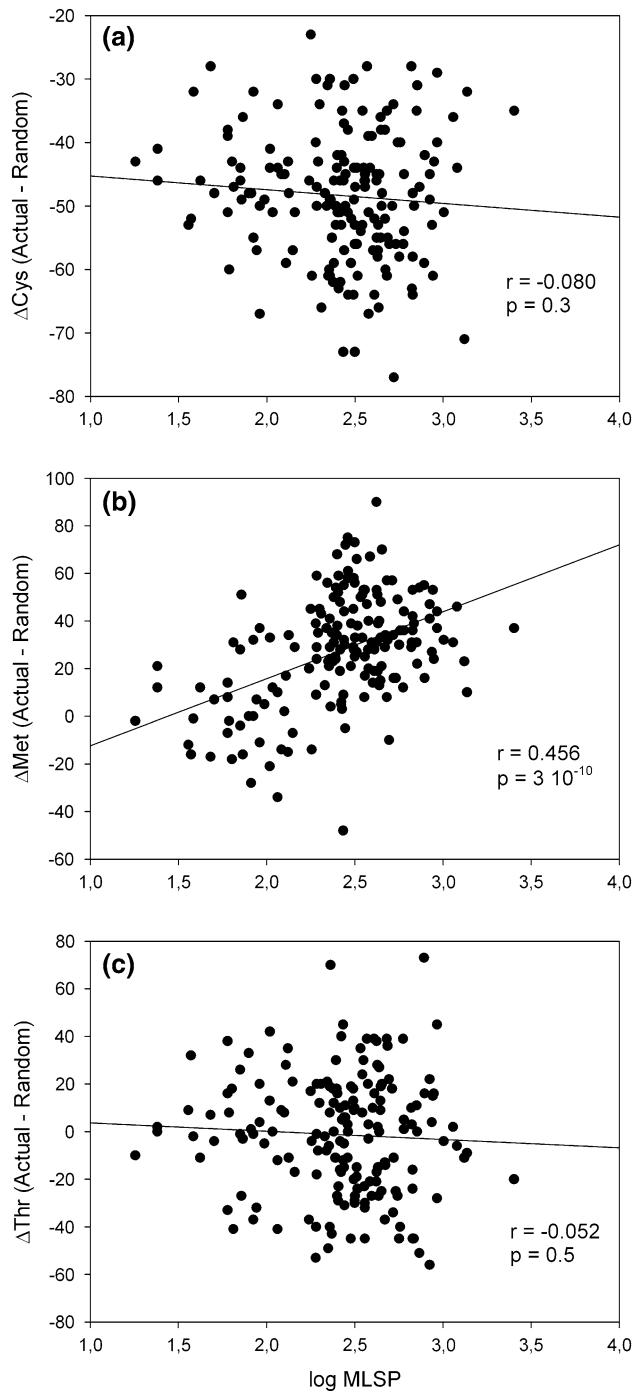


Fig. 5 Long-lived mammals actively add methionine into their mtDNA-encoded proteins. The correlations between longevity and **a** ΔCys ($r = -0.080$, p value = 0.3, $n = 173$), **b** ΔMet ($r = 0.456$, p value = 3×10^{-10} , $n = 173$) and **c** ΔThr ($r = -0.052$, p value = 0.5, $n = 173$) were analyzed

which is not possible any further lifespan-dependent cysteine depletion.

With respect to the highly significant correlation observed between ΔMet and longevity, this result is in agreement with the following hypothesis. If methionyl residues fulfill an

anti-oxidant role, then increased methionine usages may confer an adaptive advantage for both short- and long-lived species. In this context, short-lived mammals that achieve high methionine frequencies through biased mutational processes do not experience the need to depart from randomness in the same extent as long-lived animals do.

In any event, the current results challenge the view of an active selection against methionine in long-lived animals. Previous authors have pointed out a pro-oxidant role for methionyl residues (Ruiz et al. 2005). They argue that the sensitivity of proteins to oxidative stress may increase as a function of the number of methionine residues and, consequently, a lower abundance of this amino acid in proteins from long-lived animals likely contributes to the superior longevity of these species. However, our results suggest that methionine residues are not selected against in the mitochondrial proteins of long-lived mammals; on the contrary, long-lived animals seem to actively add methionine residues into their mtDNA-encoded proteins. Furthermore, we have observed a positive relationship among longevity and the departure from random methionine content (Fig. 5b). It is important to realize that this gain in methionine residues, preferentially observed in long-lived animals, is relative to random sequences with the same nucleotide composition as the actual sequences. In fact, when methionine-adding events are analyzed regardless of the nucleotide composition of the coding genes, short-lived animals exhibit higher numbers of additions than their long-lived counterparts (Aledo et al. 2011). Thus, an integrated, though somewhat speculative, interpretation of all these data may be as follows.

Because of the anti-oxidant role of methionine, both short- and long-lived mammals may benefit from incorporating this amino acid into their proteins. However, the strategies followed to achieve increased methionine usages can differ from one species to other. In this respect, short-lived mammals exhibit biases in nucleotide composition of their genes favoring the incorporation of methionine into their mitochondrial proteins (Fig. 2b). Whether these favorable biases are driven by shorter generation times (de Magalhães et al. 2007) or higher mutation rates (Nabholz et al. 2008) with respect to their long-lived counterparts are, however, issues that remain unresolved. On the other hand, long-lived mammals possess mitochondrial genes with nucleotide abundances less prone to form methionine codons when randomly reorganized into triplets (Fig. 5). Thus, if incorporating methionine into mtDNA-encoded proteins has any beneficial effect on fighting oxidative stress, this group of long-lived animals may rely more heavily on post mutational mechanisms to increase methionine usages, and therefore departure from random expectation (Fig. 5b).

In a previous work, Jobson et al. (2010) failed to find a significant correlation between cysteine, methionine, and

threonine composition and lifespan in 40 nDNA-encoded OXPHOS proteins. Herein, we have carried out the randomization analyses described above, using 61 nDNA-encoded OXPHOS genes from a dozen of mammalian species (Supplementary Table 1). As it can be deduced from Supplementary Fig. 1, while $\Delta\text{Met} > 0$ for all the analyzed species, ΔCys and ΔThr were negative in all the cases. These results may suggest that also in nDNA-encoded OXPHOS proteins methionyl residues are favored, while cysteine is avoided. However, we found no significant correlation between ΔAa and lifespan, regardless of the residue being considered. This lack of relationship between lifespan and ΔMet in nDNA-encoded proteins may be seen as a circumstantial evidence against a role of adaptation in shaping methionine content in mtDNA-encoded proteins. However, a note of caution should be made to this respect: because mtDNA-encoded proteins often evolve under different selective constraints to those of nDNA-encoded proteins. This point is well illustrated by the observation that, while nDNA-encoded residues in the interface of OXPHOS protein complexes are highly constrained, their mtDNA-encoded counterparts evolve even faster than other mtDNA-encoded residues (Schmidt et al. 2001).

In summary, methionine content in mtDNA-encoded proteins seems to be shaped by the contribution of two factors. On one hand, nucleotide composition bias brought about by a directional mutation bias is a significant driving force in methionine abundance. On the other hand, most species carry a higher number of methionyl residues into their proteins, with respect to that expected from the influence of nucleotide bias, suggesting the existence of selective forces favoring such outcome. Nevertheless, the circumstantiality of the evidences advises for caution.

Although there is a wide consensus that directional mutation bias may be related to the particular mode of replication of the mitochondrial chromosomes (Nikolaou and Almirantis 2006; Fonseca et al. 2008), the causes leading to the remarkable inter-specific difference between mitochondrial genomes are largely unknown. Thus, the question of whether the mechanisms underlying this directional mutation bias are selectively neutral or not, remains an open issue that will provide a future challenge for molecular evolutionary biologists.

Conclusion

We have presented a comparative study of mitochondrial genomes across multiple species, aiming to address the long-standing question of whether the link between amino acid usages and longevity is due to an adaptive response and/or nucleotide mutation bias. We found that the nucleotide composition bias is the main, if not the only, driving

force in threonine composition of the encoded proteins. In contrast, nucleotide composition has no effect at all on the cysteine usages, which seem to be kept at low values by purifying selection. With respect to methionine, the results suggest that both nucleotide bias and selective forces unrelated to the nucleotide composition, contribute to shape the methionine content in mtDNA-encoded proteins. Overall, our results demonstrate a role of selection in determining the composition of cysteine and methionine in the mitochondrial genome of mammals. Whether there is or not a link between the content of these sulfur-containing amino acids and the protection against oxidative stress, is an issue that remains open and deserves further attention.

Acknowledgments Thanks to Miguel Ángel Medina and Alicia Esteban del Valle for critical reading of the manuscript. We would also like to thank Richard W. Jobson for assistance in collecting data related to nDNA-encoded genes. JCA gratefully acknowledges the support of Grant CGL2010-18124 from the Ministerio de Ciencia e Innovación, Spain. JPM is supported by the BBSRC, the Wellcome Trust, the Ellison Medical Foundation and a Marie Curie International Reintegration Grant within EC-FP7.

Conflict of interest The authors declare that they have no conflict of interest.

References

- Albu M, Min XJ, Hickey D, Golding B (2008) Uncorrected nucleotide bias in mtDNA can mimic the effects of positive Darwinian selection. *Mol Biol Evol* 25:2521–2524
- Aledo JC, Li Y, de Magalhães JP, Ruíz-Camacho M, Pérez-Claros JA (2011) Mitochondrially encoded methionine is inversely related to longevity in mammals. *Aging Cell* 10:198–207
- Balaban RS, Nemoto S, Finkel T (2005) Mitochondria, oxidants, and aging. *Cell* 120:483–495
- Barja G (1998) Mitochondrial free radical production and aging in mammals and birds. *Ann N Y Acad Sci* 854:224–238
- Bender A, Hajieva P, Moosmann B (2008) Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. *Proc Natl Acad Sci USA* 105:16496–16501
- Berlett BS, Stadtman ER (1997) Protein oxidation in aging, disease, and oxidative stress. *J Biol Chem* 272:20313–20316
- Brand MD (2000) Uncoupling to survive? The role of mitochondrial inefficiency in ageing. *Exp Gerontol* 35:811–820
- Collins D, Jukes T (1993) Relationship between G + C in silent sites of codons and amino acid composition of human proteins. *J Mol Evol* 36:201–213
- D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol* 32:504–510
- de Magalhães JP, Costa J (2009) A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol* 22:1770–1774
- de Magalhães JP, Costa J, Church GM (2007) An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J Gerontol A* 62:149–160
- Fonseca MM, Posada D, Harris DJ (2008) Inverted replication of vertebrate mitochondria. *Mol Biol Evol* 25:805–808
- Foster PG, Jermini LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44:282–288
- Harman D (2006) Free radical theory of aging: an update: increasing the functional life span. *Ann N Y Acad Sci* 1067:10–21
- Jia W, Higgs PG (2008) Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol Biol Evol* 25:339–351
- Jobson RW, Dehne-Garcia A, Galtier N (2010) Apparent longevity-related adaptation of mitochondrial amino acid content is due to nucleotide compositional shifts. *Mitochondrion* 10:540–547
- Kitazoe Y, Kishino H, Hasegawa M, Nakajima N, Thorne JL, Tanaka M (2008) Adaptive threonine increase in transmembrane regions of mitochondrial proteins in higher primates. *PLoS ONE* 3:e3343
- Kitazoe Y, Kishino H, Hasegawa M, Matsui A, Lane N, Tanaka M (2011) Stability of mitochondrial membrane proteins in terrestrial vertebrates predicts aerobic capacity and longevity. *Genome Biol Evol* 3:1233–1244
- Lambert AJ, Boysen HM, Buckingham JA, Yang T, Podlutsky A, Austad SN, Kunz TH, Buffenstein R, Brand MD (2007) Low rates of hydrogen peroxide production by isolated heart mitochondria associate with long maximum lifespan in vertebrate homeotherms. *Aging Cell* 6:607–618
- Lambert AJ, Buckingham JA, Boysen HM, Brand MD (2010) Low complex I content explains the low hydrogen peroxide production rate of heart mitochondria from the long-lived pigeon, *Columba livia*. *Aging Cell* 9:78–91
- Levine RL, Mosoni L, Berlett BS, Stadtman ER (1996) Methionine residues as endogenous antioxidants in proteins. *Proc Natl Acad Sci USA* 93:15036–15040
- Min XJ, Hickey DA (2007) DNA asymmetric strand bias affects the amino acid composition of mitochondrial proteins. *DNA Res* 14:201–206
- Moosmann B, Behl C (2008) Mitochondrially encoded cysteine predicts animal lifespan. *Aging Cell* 7:32–46
- Nabholz B, Glémis S, Galtier N (2008) Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. *Mol Biol Evol* 25:120–130
- Nikolaou C, Almirantis Y (2006) Deviations from Chargaff's second parity rule in organellar DNA insights into the evolution of organellar genomes. *Gene* 381:34–41
- Pamplona R (2011) Mitochondrial DNA damage and animal longevity: insights from comparative studies. *J Aging Res* 2011:807108
- Pamplona R, Barja G (2011) An evolutionary comparative scan for longevity-related oxidative stress resistance mechanisms in homeotherms. *Biogerontology* 12:409–435
- Pamplona R, Costantini D (2011) Molecular and structural antioxidant defenses against oxidative stress in animals. *AJP Regul Integr Comp Physiol* 301:R843–R863
- Pamplona R, Portero-Otín M, Riba D, Ruiz C, Prat J, Bellmunt MJ, Barja G (1998) Mitochondrial membrane peroxidizability index is inversely related to maximum life span in mammals. *J Lipid Res* 39:1989–1994
- Pamplona R, Portero-Otín M, Sanz A, Ayala V, Vasileva E, Barja G (2005) Protein and lipid oxidative damage and complex I content are lower in the brain of budgerigars and canaries than in mice. Relation to aging rate. *Age* 27:267–280
- Ruiz MC, Ayala V, Portero-Otín M, Requena JR, Barja G, Pamplona R (2005) Protein methionine content and MDA-lysine adducts are inversely related to maximum life span in the heart of mammals. *Mech Ageing Dev* 126:1106–1114

- Samuels DC (2005) Life span is related to the free energy of mitochondrial DNA. *Mech Ageing Dev* 126:1123–1129
- Schmidt TR, Wu W, Goodman M, Grossman LI (2001) Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Mol Biol Evol* 18:563–569
- Singer GA, Hickey DA (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* 17:1581–1588
- Speakman JR, Talbot DA, Selman C, Snart S, McLaren JS, Redman P, Krol E, Jackson DM, Johnson MS, Brand MD (2004) Uncoupled and surviving: individual mice with high metabolism have greater mitochondrial uncoupling and live longer. *Aging Cell* 3:87–95
- Stadtman ER, Van Remmen H, Richardson A, Wehr NB, Levine RL (2005) Methionine oxidation and aging. *Biochim Biophys Acta* 1703:135–140
- Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harbor symposium on quantitative biology*
- Wang H-C, Singer GAC, Hickey DA (2004) Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol* 21:90–96

E.2 Publicación del capítulo 2

Thermodynamic Stability Explains the Differential Evolutionary Dynamics of Cytochrome b and COX I in Mammals

Juan Carlos Aledo · Héctor Valverde ·
Manuel Ruiz-Camacho

Received: 11 September 2011/Accepted: 2 February 2012/Published online: 24 February 2012
© Springer Science+Business Media, LLC 2012

Abstract By using a combination of evolutionary and structural data from 231 species, we have addressed the relationship between evolution and structural features of cytochrome b and COX I, two mtDNA-encoded proteins. The interior of cytochrome b, in contrast to that of COX I, exhibits a remarkable tolerance to changes. The higher evolvability of cytochrome b contrasts with the lower rate of synonymous substitutions of its gene when compared to that of COX I, suggesting that the latter is subjected to a stronger purifying selection. We present evidences that the stability effect of mutations ($\Delta\Delta G$) may be behind these differential behaviour.

Keywords COX I · Cytochrome b · Evolvability · mtDNA · Natural selection · Protein evolution

Introduction

In addition to their central role in the oxidative phosphorylation (OXPHOS), mitochondria are involved in many cellular processes such as growth, apoptosis and ageing (Aledo 2004; Aledo et al. 2011; Navarro and Boveris 2007). Not surprisingly, mitochondrial defects have been

associated with a number of diseases (Scharfe et al. 2009), which may be the result of spontaneous or inherited mutations in the mitochondrial genome (mtDNA) or in nuclear genes (nDNA) that code for mitochondrial components (DiMauro and Schon 2008; Gallardo et al. 2006). In mammals, mtDNA encodes only 13 proteins of the respiratory chain, while the bulk of mitochondrial proteins are encoded by nuclear genes. The evolution of mtDNA contrasts with that of nDNA. Indeed, mitochondrial and nuclear genomes differ in many ways, such as the total length, ploidy level, mode of inheritance, recombination rate, presence of introns, effective population size and repair mechanisms. The fact that mtDNA has been evolving much more rapidly than nDNA in higher animals is currently undisputed, although it came as a surprise (Brown et al. 1979). Given the importance of the mtDNA-encoded proteins in OXPHOS, their more rapid rates of change seemed to challenge the idea that the more important the function of a protein, the more slowly it undergoes evolutionary change in primary structure (Zhang and He 2005).

Identifying factors that determine protein evolution rate has attracted considerable attention. Recent evidence suggests that all the events influencing protein expression, such as transcription initiation, splicing and translation, need to be considered when explaining variation in the rate at which different proteins evolve (Pál et al. 2006; Warnecke et al. 2009). However, when addressing the evolutionary rate of different residues within a single protein, the attention should be focused on functional–structural aspects (Franzosa and Xia 2009). In this sense, the functional properties of a protein, including the interactions with other proteins and post-translational modifications, are all related to its surface properties. However, the ability of this protein to fold correctly and the thermodynamic stability of this fold, which ultimately assures the protein function, are greatly

Electronic supplementary material The online version of this article (doi:[10.1007/s00239-012-9489-0](https://doi.org/10.1007/s00239-012-9489-0)) contains supplementary material, which is available to authorized users.

J. C. Aledo (✉) · H. Valverde
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain
e-mail: caledo@uma.es

M. Ruiz-Camacho
Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain

influenced by features of the protein interior (Eilers et al. 2000; Mirny and Shakhnovich 2001). Therefore, exposure to solvent is a structural property that has received particular attention as a potential determinant of protein evolution.

Studies carried out on yeast (Conant and Stadler 2009; Franzosa and Xia 2009; Lin et al. 2007) and bacterial proteins (Bustamante et al. 2000) support the view that residues buried in a protein's core are most likely to remain conserved during evolution compared to their solvent exposed counterparts. From these findings, one may be tempted to speculate that proteins with a small proportion of solvent exposed residues should evolve slowly. Although such reasoning has received support from some authors (Lin et al. 2007), others have reported the opposite observation, that is, proteins with fewer exposed residues evolve more rapidly (Bloom et al. 2006a). In a recent work, Franzosa and Xia (2009) suggested that increasing core size has little effect on evolutionary rate among solvent excluded residues while yields a more rapid relaxation of constraint for those residues exposed to the solvent.

Since the effect of structural features on protein evolution has been matter of debate as to its mechanism and significance, we found interesting to address these issues using mtDNA-encoded proteins, which often show a differential evolutionary pattern with respect to nuclear-encoded proteins (Schmidt et al. 2001; Welch et al. 2008). In the current work, we have explored the evolutionary dynamics of the interior and surface of cytochrome b and COX I, two mtDNA-encoded proteins commonly used in phylogenetic studies.

Materials and Methods

Data Sources and Molecular Modelling

A collection of 231 mammalian mitochondrial genomes (Fig. 1) was obtained from the National Center for Biotechnology Information (NCBI) genome database (www.ncbi.nlm.nih.gov). These mammalian species encompass 27 orders: Afrosoricida ($n = 3$), Artiodactyla ($n = 25$), Carnivora ($n = 56$), Cetacea ($n = 24$), Chiroptera ($n = 10$), Cingulata ($n = 1$), Dasyuromorphia ($n = 5$), Dermoptera ($n = 1$), Delphimorphia ($n = 4$), Diprotontiae ($n = 11$), Erinaceomorpha ($n = 4$), Hyracoidea ($n = 2$), Lagomorpha ($n = 5$), Macroscelidea ($n = 2$), Monotremata ($n = 3$), Paramelemorphia ($n = 4$), Paucituberculata ($n = 2$), Perissodactyla ($n = 5$), Pholidota ($n = 1$), Pilosa ($n = 3$), Primates ($n = 32$), Proboscidea ($n = 4$), Rodentia ($n = 13$), Scandentia ($n = 1$), Sirenia ($n = 2$), Soricomorpha ($n = 7$) and Tubulidentata ($n = 1$). Multiple sequence alignments of orthologous proteins were performed using ClustalX 2.0.9. The gap opening and extension penalties were 15 and 6.66,

respectively. The delay divergent sequences option was set to 30%. A value of 0.5 was chosen for the transition weight parameter. Sequence identities were higher than 73 and 90% for cytochrome b and COX I, respectively. Three-dimensional models structures for cytochrome b and COX I were generated by alignment with the experimental crystal structures of corresponding bovine sequences (Protein Data Bank, PDB, 1be3 chain C and 2occ chain A, respectively). Structural calculations were performed on the Swiss-Model workspace (Bordoli et al. 2009). Owing to the difficulties of obtaining reliable structural models for all the 231 mammalian sequences, the analyses described below were done using sets containing 221 and 189 PDB files for cytochrome b and COX I, respectively (Fig. 1).

Determining Exposed and Buried Positions

Solvent accessible surface areas (ASA) were computed using the SurfRace program (Tsodikov et al. 2002). The accessibility of a given amino acid residue in a protein was calculated as the ratio of its ASA in the native protein structure to that it would have in an unfolded and extended polypeptide chain ($\Phi = -120^\circ$, $\Psi = 140^\circ$), with the side-chain conformations corresponding to the one most frequently observed in proteins (Miller et al. 1987). Amino acid residues exhibiting accessibilities below 5% were defined as buried residues. Then, following a multiple sequence alignment of orthologous proteins, the number of instances (species) a given position appeared as buried was computed. This position was considered as a buried position when in most of the species (>50%) the residue found showed an accessibility below the threshold of 5%.

Shannon's Entropy Determinations

For the mitochondrial protein being analyzed, multiple alignments of orthologous sequences were obtained, which allowed us to compute variability at position j of the alignment as the corresponding Shannon's entropy:

$$H(j)_c = - \sum_{i=1}^c p_i(j) \log_c p_i(j) \quad (1)$$

where $p_i(j)$ is the frequency of residue from class i in position j . Since different amino acid classification schemes can be contemplated, c stands for the number of different classes. More concretely, we used six classes of residues, $c = 6$, to reflect physico-chemical properties of amino acids and their natural pattern of substitution (Mirny and Shakhnovich 2001; Thompson and Goldstein 1996). In addition, for each position the Shannon's entropy was computed without grouping amino acids, that is, for $c = 20$.



Fig. 1 Mammalian species used in the study. A collection of 231 mammalian mitochondrial genomes was obtained from the National Center for Biotechnology Information. For each species, the amino acid sequences of cytochrome b and COX I were concatenated and subjected to alignments to construct a tree using the program Promlk from the PHYLIP package. The name of each species is accompanied

A low entropy value for a given position is interpreted as a lowly variable position. The extreme case being an invariant position, which would yield a null entropy. In the opposite extreme, an entropy value of 1 for a given position, is interpreted as a highly variable position.

by a bidimensional vector. The first coordinate of the vector is 1 if a reliable structural model for cytochrome b could be obtained for that species, and 0 in the opposite case. Similarly, the second coordinate is 1 or 0, depending on whether or not a COX I structural model could be obtained, respectively

Determining Unconstrained, Conservative and Radical Positions

The comparison between $H_6(j)$ and $H_{20}(j)$ can provide some insights into the forces behind the evolution of

residues at position j . For instance, positions yielding low values for both entropy measures, H_6 and H_{20} , can be considered as constrained positions. In contrast, those positions exhibiting high values for both entropy functions are expected to be less critical, allowing a relaxation of the purifying selection. These positions are designated as *unconstrained*. On the other hand, a high value for H_{20} along with a low H_6 value inform us that at that position, the properties of the residues rather than the amino acids themselves, are critical and subjected to selection; thus, we labelled this positions as *conservative*. More interesting are those positions exhibiting a low H_{20} value accompanied by high H_6 entropy. In this case, we have few changes at that positions (low H_{20}) but these changes are radical because they are taking place between amino acids belonging to groups with very different properties (high H_6), which may reflect adaptive changes driven by positive selection. These positions are referred to as *radical*. Thus, for each protein we assembled three sets of positions designed as *Unconstrained*, *Conservative* and *Radical*, according to the following criteria:

$$\text{Unconstrained} = (j \in J : H_{20}(j) \geq \text{UQ}_{20} \text{ and } H_6(j) \geq \text{UQ}_6)$$

$$\text{Conservative} = (j \in J : H_{20}(j) \geq \text{UQ}_{20} \text{ and } H_6(j) \leq \text{LQ}_6)$$

$$\text{Radical} = (j \in J : H_{20}(j) \leq \text{LQ}_{20} \text{ and } H_6(j) \geq \text{UQ}_6)$$

where J is the set containing either the positions of the 262 variable sites of cytochrome b or the 230 variable positions of COX I. LQ_c and UQ_c are the lower and upper quartiles of Shannon's entropy distribution, respectively. To calculate these quartiles, invariant positions, $H_{20}(j) = 0$, were previously removed. In this way, for cytochrome b we obtained the following values: $\text{LQ}_{20} = 0.0079$ and $\text{UQ}_{20} = 0.0501$; $\text{LQ}_6 = 0.0049$ and $\text{UQ}_6 = 0.0650$. For COX I, the quartiles took the following values: $\text{LQ}_{20} = 0.008$, $\text{UQ}_{20} = 0.0376$, $\text{LQ}_6 = 0.0040$ and $\text{UQ}_6 = 0.0222$.

mtDNA Substitutions Rates

Whole-tree estimation can be unreliable if nuisance parameters, such as base composition, vary across groups. This problem is particularly relevant to mitochondrial genes, since uncorrected nucleotide bias in mtDNA can mimic the effect of positive selection (Albu et al. 2008). For these reasons, we only calculated substitution rates for closely related pairs of species (see Online Resource 1). To assist in the assemblage of a data set formed by pairs of close species, we generated maximum likelihood phylogenetic subtrees using the PHYLIP package. Each subtree accounted for the species belonging to the same mammalian order. Only those species directly connected to the same internal node were

considered as a suitable pair. A phylogeny reliably reconstructed from data unrelated to the primary structure of cytochrome b and COX I was preferable to avoid any semblance of circularity. Thus, the complete sequences of the two mitochondrial ribosomal RNA genes were concatenated for each species, and used to generate the phylogenetic reconstructions that assisted in the assemblage of an initial data set formed by 54 pairs of close species. We then computed the mean number of nucleotide differences per site by pairwise comparison using the Nei–Gojobori method (Nei and Gojobori 1986) and the Jukes–Cantor correction to account for multiple substitutions at the same site. In this way, for each gene (cytochrome b and COX I) we obtained 54 points of the $d_S \times d_N$ plane, where the d_S variable represents the mean number of synonymous differences per synonymous site while d_N is the mean number of nonsynonymous differences per nonsynonymous site.

Next, each DNA orthologous sequence was divided into two sets on the basis of the accessibility of the amino acid residue being encoded for the considered codon. In other words, mtDNA nucleotide sequences encoding cytochrome b and COX I buried residues were segregated and placed in a separate data set from those encoding for exposed residues. Afterwards, for each data set d_S and d_N were computed as explained above.

Thermodynamic Stability Changes

The thermodynamic stability changes ($\Delta\Delta G$) of mutations were computed using the protein design tool FoldX version 3.0 (Guerois et al. 2002; Schymkowitz et al. 2005; Tokuriki et al. 2007). FoldX uses a full atomic description of the structure of the protein, to provide a quantitative estimation of the importance of the interactions contributing to the stability of this protein. The different energy terms taken into account, which have been described in detail somewhere else (Guerois et al. 2002), have been weighted using empirical data obtained from protein engineering experiments.

3D structures for both cytochrome b and COX I were subjected to an optimization procedure using the repair function of FoldX. Then for each protein in each species, an alanine scan was carried out. That is, every single residue was replaced by alanine one by one, and the resulting $\Delta\Delta G$ was computed and recorded as a function of the residue position in the primary protein structure. This procedure provided two matrices of 221×379 and 189×514 , containing $\Delta\Delta G$ values for cytochrome b and COX I, respectively (raw data can be provided under request).

Computation and Statistical Analyses

Random distributions were generated using Perl scripts. Probability calculations were assisted by Wolfram

Mathematica 8.0. All other statistic analyses were done with SPSS 15.0.

Results

Invariant Positions are Accumulated in the Interior of COX I but May be Randomly Distributed Through the Whole Cytochrome b Protein

To address whether or not the surface and interior of mitochondrial proteins evolve differentially, we started sorting out each residue position as buried or surface site according to the criteria exposed in the methodological section. Once this segregation was accomplished, we addressed whether evolutionary rates varied between these different residue sets. As a first approach, we focused our interest on those residues that have remained invariant during the diversification of mammals. More concretely, we tested the following null hypothesis: invariant residues are randomly distributed between the interior and surface of the considered proteins. To this end, we defined the random variable X as the number of invariant residues that are buried in the protein interior. Beside the current value of X (designed by lowercase x), the proportion of invariant residues ($p_x = \text{number of invariant residues/total number of residues}$) and the number of buried residues, n , were computed for each protein (Table 1). In this way, under the null hypothesis conditions, we can assume that X follows a binomial distribution, $X \sim \text{Bin}(n, p_x)$, which allows us to calculate the probability of finding by chance a number of invariant buried residues equal or higher to that observed for each protein, $P[X \geq x]$. Although such probabilities were relatively low for both proteins cytochrome b and COX I, only in the latter case the null hypothesis could be rejected at a significance level of 1%.

COX I, but Not Cytochrome b, Protein Interior Shows a Low Shannon's Entropy that Departs from Random Expectations

Hitherto we have analyzed the departure from random distribution of invariant positions, considering a position as

Table 1 Abundances of exposed, buried and invariant residues in cytochrome b and COX I

Protein	P_{exp}	p_x	n	x	$P[X \geq x]$
Cytochrome b	0.8482	0.314	58	22	0.175
COX I	0.6718	0.556	170	110	0.009

P_{exp} is the proportion of exposed residues, p_x is the proportion of invariant residues, n is the number of buried residues, x is the number of invariant buried residues, $P[X \geq x]$ is the probability of finding by chance a number of invariant buried residues equal to or higher than that observed

invariant when the same amino acid is found, without exception, in all the analyzed species. We next used an information theoretic formalism to study the evolutionary conservation of the protein interiors.

After computing H_6 and H_{20} for each position, the averaged entropies for the buried residues were worked out. The mean values for $H_6(\text{buried})$ and $H_{20}(\text{buried})$ were 0.023 and 0.019, respectively, in the case of cytochrome b. These values went down to $H_6(\text{buried}) = 0.006$ and $H_{20}(\text{buried}) = 0.005$, when COX I was considered. In this way, the conservation of the buried protein core can be statistically compared with the conservation of all residues in the protein. To this end, we tested the following null hypothesis: buried residues are not more conserved than the whole protein sequence. To contrast this hypothesis, we compared the above means with the distribution of mean entropy values of the same number of residues randomly chosen from the whole protein, $H_c(\text{random})$. These distributions were obtained by taking 10^5 random sets of 58 residues for cytochrome b and 170 residues for COX I. Then, the fraction of instances with $H_c(\text{random}) < H_c(\text{buried})$ gives the probability of observing by chance a mean entropy value lower than that computed for the buried residues. In other words, it gives us the type I error rate. We were unable to reject the null hypothesis for cytochrome b at a confidence level of $\alpha = 10\%$. In contrast, for COX I the hypothesis was rejected at confidence levels as low as $\alpha = 0.34$ and 0.00%, for H_6 and H_{20} , respectively. Thus, these results are in line with those previously shown on invariant positions.

Since COX I buried positions are preferentially enriched with invariant residues, it might be that the low $H_c(\text{buried})$ we have reported above were due to the contribution of invariant amino acids from the protein core. In other words, we wanted to address whether buried positions were still more constrained than positions at the protein surface, once those invariant positions were excluded from the study. For this purpose, random sets of 36 and 60 residues from cytochrome b and COX I, respectively, were used to generate random distributions as explained above. Figure 2 shows the results of such analyses. As it can be deduced from this figure, the variable buried residues from COX I are much more constrained than the rest of COX I variable positions ($\alpha \leq 5\%$). However, variable buried positions from cytochrome b failed again to exhibit a statistically lower variability with respect to their exposed counterpart ($\alpha > 18\%$).

Unconstrained Positions are Randomly Distributed in Cytochrome b but are Selectively Excluded from the Interior of COX I

For each protein we assembled three sets of positions designed as *Unconstrained*, *Conservative* and *Radical*,

according to the criteria specified in “[Determining Unconstrained, Conservative and Radical Positions](#)” (Fig. 3). Afterwards, we assessed whether or not these position categories were preferentially located in or excluded from the protein interior. To this end, we computed the frequencies of unconstrained ($p_u = 0.068$ and 0.054), conservative ($p_c = 0.018$ and 0.037) and radical position ($p_r = 0.034$ and 0.031) for cytochrome b and COX I, respectively. These frequencies were used as proxies of the probabilities of a given position being an unconstrained, conservative or radical position, respectively. Also, we defined three random variables (U , C , R) as the number of unconstrained buried, conservative buried and radical buried positions. If these three types of positions are randomly distributed between the protein interior and the protein surface, one would expect that each of these random variables would follow a binomial distribution: $U \sim Bi(n, p_u)$, $C \sim Bi(n, p_c)$, $R \sim Bi(n, p_r)$; where n is the number of buried residues of the protein under consideration. In this way, after computing the current values that these random variables take for each protein (represented by lower-case letters: u , c and r , respectively), we were in conditions to calculate the probability of finding by chance, in the protein interior, a number equal or higher to that observed. That is, $P[U \geq u]$, $P[C \geq c]$ and $P[R \geq r]$. Since all these probabilities were much higher than 0.01, for both cytochrome b and COX I (Table 2), we concluded that the

interior of these mitochondrial proteins is not particularly enriched in any of these type of positions. We next tested the possibility of any of these position categories being selectively excluded from the protein core. For this purpose, we calculated, under the null hypothesis conditions (random distribution), the probability of finding by chance, in the protein interior, a number of unconstrained, conservative or radical positions lower or equal to that observed for each protein. That is, $P[U \leq u]$, $P[C \leq c]$ and $P[R \leq r]$. When these probabilities are below 0.01 we can reject the null hypothesis of random distribution. As it can be observed in Table 2, only COX I gave a $P[U \leq u]$ value below the threshold of 0.01, which suggests that unconstrained positions are selectively excluded from the COX I core, but not from the cytochrome b interior.

The COX I Gene is Subjected to Both a Higher Mutation Pressure and a Stronger Purifying Selection than the Cytochrome b Gene

Under the assumption that changes at silent sites are mainly neutral (Kimura 1977), the comparison between the rates of synonymous substitutions per site within the cytochrome b with those for COX I, suggested a surprisingly higher mutational pressure for COX I with respect to cytochrome b (Fig. 4a). In contrast, when the rates of nonsynonymous substitutions per site were the subject of comparisons, COXI

Fig. 2 Shannon's entropy of cytochrome b and COX I buried positions. After removing invariant positions, the mean H_6 and H_{20} entropies for buried residues were calculated and are shown in the upper-left corner of each plot. These mean values were compared with the distribution of mean entropy values of the same number of residues randomly chosen from the protein (surface + interior)

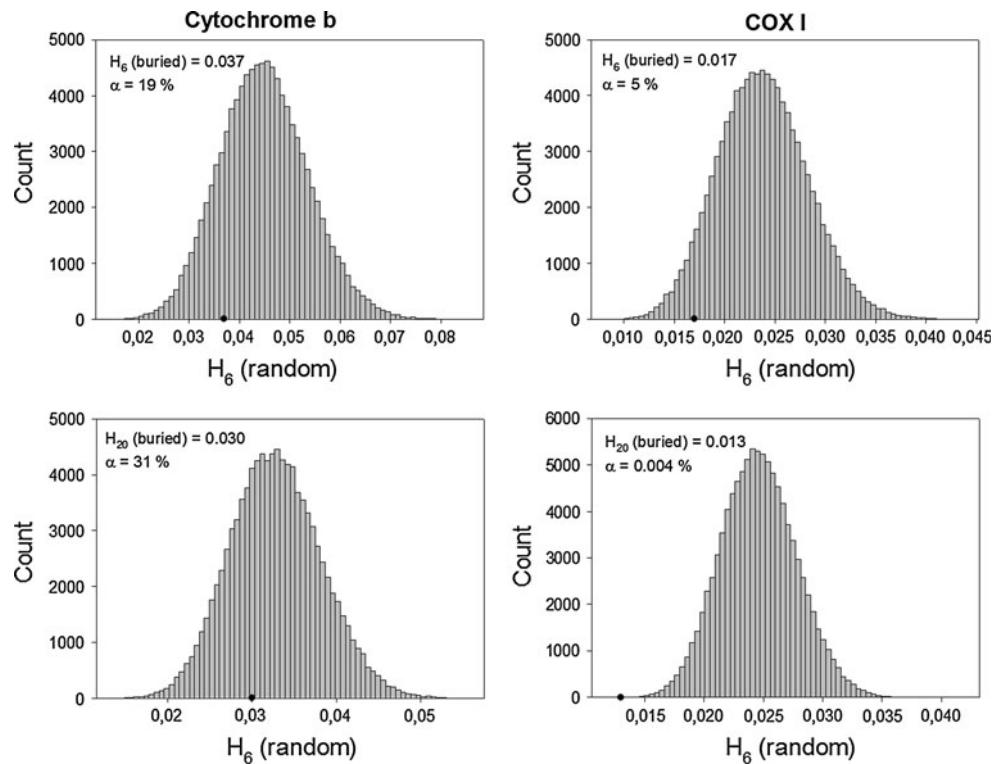
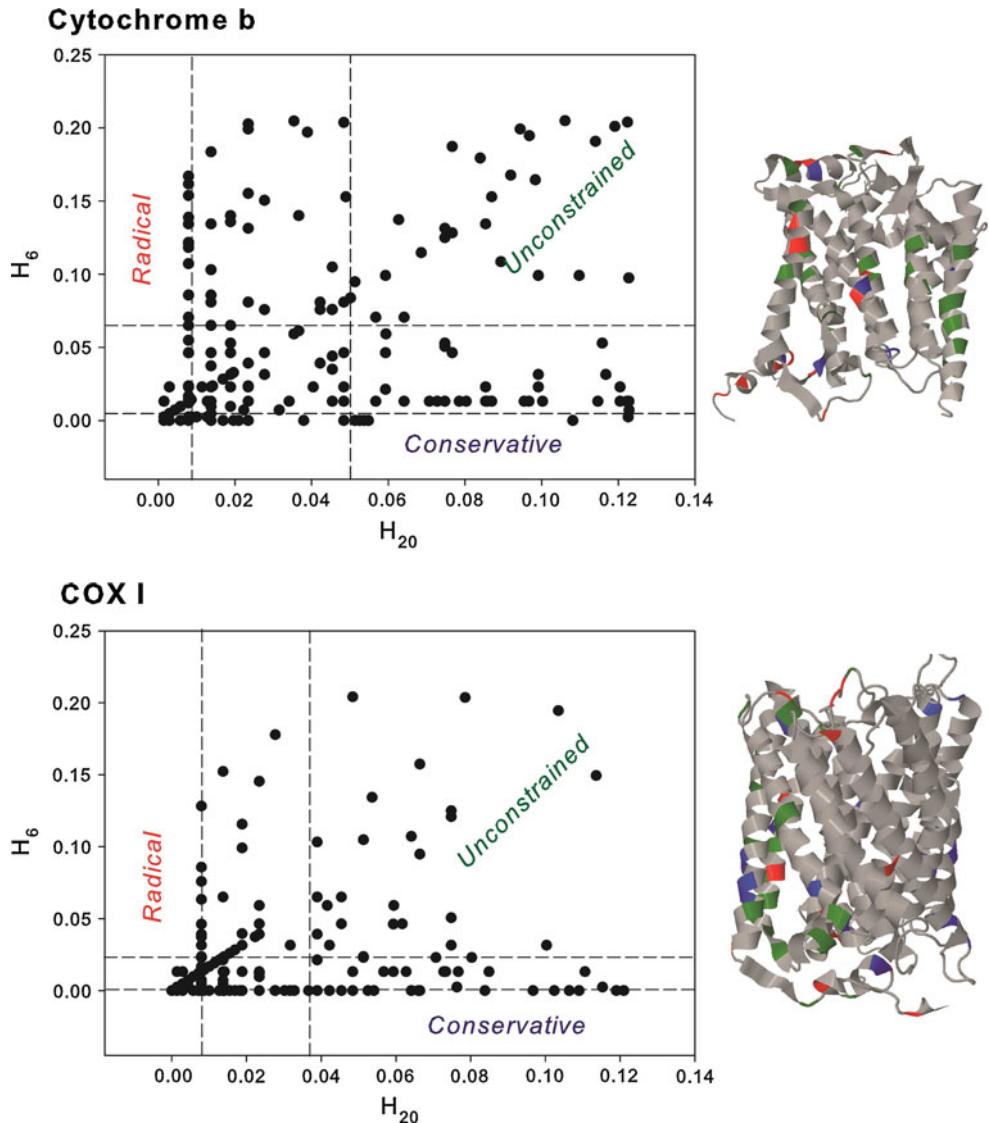


Fig. 3 Identification and location of unconstrained, conservative and radical sites. H_6 was plotted against H_{20} and horizontal dashed lines were drawn to indicate the lower and upper quartiles of the H_6 distribution. Similarly, vertical dashed lines indicate the lower and upper quartiles of the H_{20} distribution. Residues belonging to each of the three categories (see the text for details) were identified on the folded structure (either cytochrome b or COX I) according to the following colour code: unconstrained (green), conservative (blue) and radical (red)



showed d_N values significantly lower than those calculated for cytochrome b, indicating that the COX I protein is much more constrained by purifying selection (Fig. 4b).

We next addressed the question of whether the rates of mutation and selection varied between those codons coding for buried amino acids and those coding for exposed residues. As shown in Fig. 4c, d, we failed to observe significant differences in d_S between the interiors and surfaces. In contrast, the rates of nonsynonymous substitutions were remarkably lower among nucleotides encoding for buried residues, regardless the gene being considered (Fig. 4e, f). Nevertheless, d_N for buried COX I was still much lower than d_N for buried cytochrome b (Fig. 5). Of note is that the rate of nonsynonymous substitutions affecting cytochrome b buried residues was much higher than that for solvent-exposed COX I residues (Fig. 5c).

Thermodynamic Stability Can Explain the Differential Behaviour of COX I and Cytochrome b

Often a single amino acid substitution can dramatically alter the stability of a protein. Not surprisingly, protein stability has been pointed out as a determinant of evolvability (Bloom et al. 2006b; Tokuriki and Tawfik 2009). According to this view, the stability effects of mutations may underlie the differential evolutionary dynamics of cytochrome b and COX I described above. In other words, we hypothesized that mutations affecting buried residues in COX I are more destabilizing than mutations taking place in the interior of cytochrome b. To address this working hypothesis, we assessed the thermodynamic stability effect ($\Delta\Delta G$) of substituting each single buried residue into alanine. To this end, the thermodynamic stability changes of

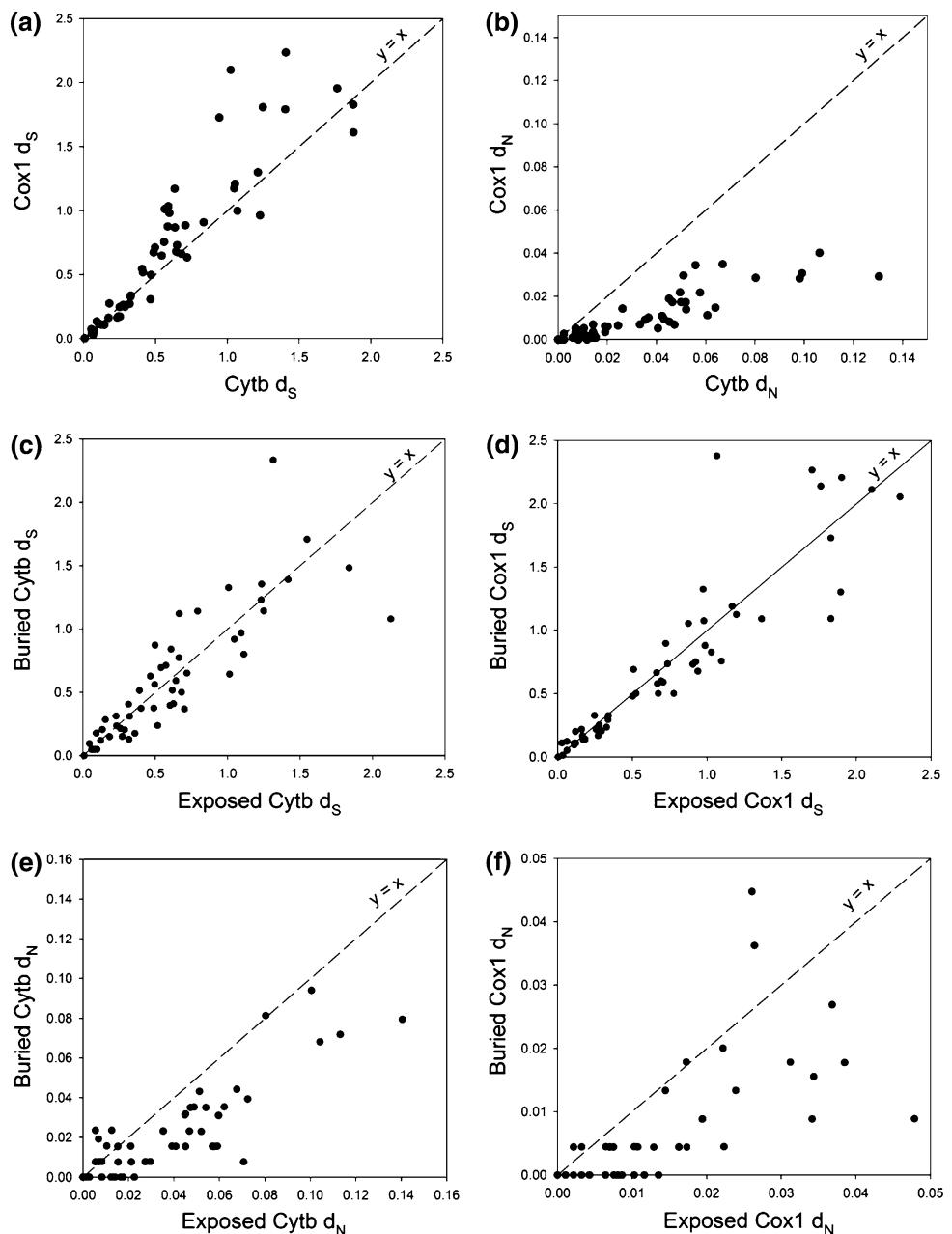
Table 2 Probabilities, according to the null hypothesis, of finding in the protein interior a number of unconstrained (u), conservative (c), or radical (r) positions lower or higher to those observed for each protein

Protein	n	$P[U \leq u]$	$P[C \leq c]$	$P[R \leq r]$	$P[U \geq u]$	$P[C \geq c]$	$P[R \geq r]$
Cyt b	58	0.438	0.913	0.135	0.764	0.281	1.000
COX I	170	0.008**	0.05	0.101	0.991	0.983	0.969

n is the number of buried residues of the protein under consideration

The null hypothesis assumes that the numbers of buried unconstrained, conservative and radical positions follow binomial distributions. See the text for details. ** Confidence level lower than 0.01

Fig. 4 Intergenic and intragenic substitution rate comparisons. For each pair of proximal species, the substitution rate per site was computed using different sets of residues from cytochrome b and COX I (see text for details), and plotted for comparative purposes



mutations were computed using the force-field FoldX (Guerois et al. 2002; Schymkowitz et al. 2005).

First, we formulated the null hypothesis that single point mutations at buried positions are no more destabilizing

than mutations at randomly chosen positions from the whole protein. To test this hypothesis we computed the mean free energy change upon mutation of buried residues to alanine, and compared it with the distribution of means

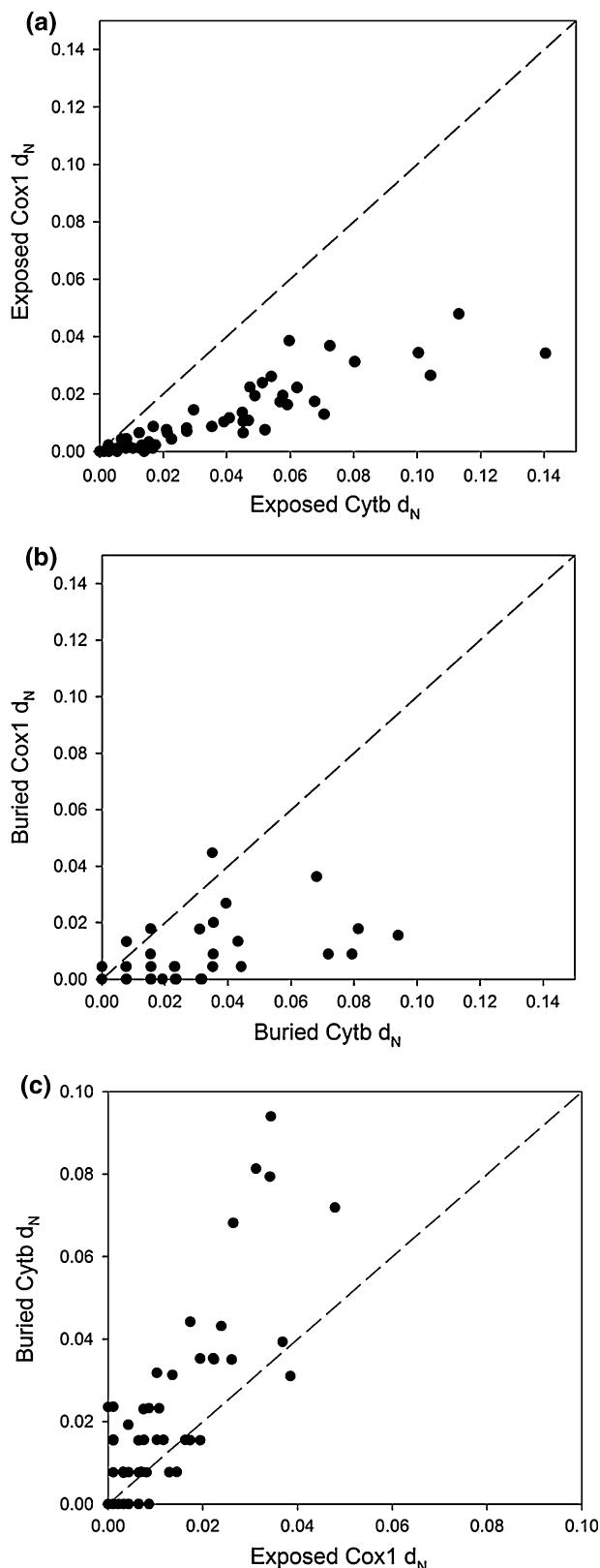


Fig. 5 Comparison of nonsynonymous substitution rates among different residues sets. The rates of nonsynonymous substitutions per site affecting the exposed regions of cytochrome b and COX I were compared (a). Similarly, the substitutions rates within the interior of these proteins were directly compared (b). The interior of cytochrome b was also compared to the surface of COX I (c)

COX I) residues and computing the mean $\Delta\Delta G$ upon single point mutation of each residue from the set. As expected, mutations affecting buried residues from COX I were strongly destabilizing. The null hypothesis could be rejected at a significance level as low as $\alpha < 10^{-6}$. Changes affecting the cytochrome b interior were also significantly ($\alpha = 0.0002$) more destabilizing than changes on the surface. However, when the $\Delta\Delta G$ distributions for cytochrome b and COX I were compared (Fig. 6), a clear-cut conclusion emerged: cytochrome b seemed to be much more robust to mutations from the thermodynamic stability point of view. Thus, nonsynonymous substitutions affecting buried residues may be easily tolerated in the case of cytochrome b, but much more unlikely for COX I, as also suggested in Fig. 5.

Both, cytochrome b and COX I are single polypeptides that form part of large multisubunit complexes. Therefore, many of the residues being classified as exposed in the single polypeptide chain, may be involved in functional and structural interactions with other subunits. Since these interactions would generate evolutionary constraints and change the stability status of new mutations, it seemed relevant to examine this issue. To this end, and using the bovine quaternary structures of complexes III and IV, we determined the $\Delta\Delta G$ after carrying out alanine scans of cytochrome b and COX I while being part of their respective complexes. Figure 6 summarizes the results of such analyses. While the $\Delta\Delta G$ distribution for cytochrome b as part of the cytochrome bc1 complex was similar to that of the single chain (with respect to mean and variance) the distribution of COX I in the complex IV showed a remarkable increased mean and reduced variance (compare Fig. 6a, b). More importantly, the comparison of $\Delta\Delta G$ for cytochrome b and COX I as part of their respective complexes, strongly suggest that, from the thermodynamic point of view, cytochrome b is much more robust to mutations than COX I.

Discussion

Understanding variability in substitution rates between different proteins and different regions of proteins is of considerable interest to molecular evolutionists, as well as to biotechnologists engaged in protein engineering. Studies carried out in bacteria (Bustamante et al. 2000) and more recently in yeast (Bloom et al. 2006a; Conant and Stadler

of the same number of residues randomly chosen in the same protein. This distribution was obtained by choosing 10^6 random sets of either 58 (for cytochrome b) or 170 (for

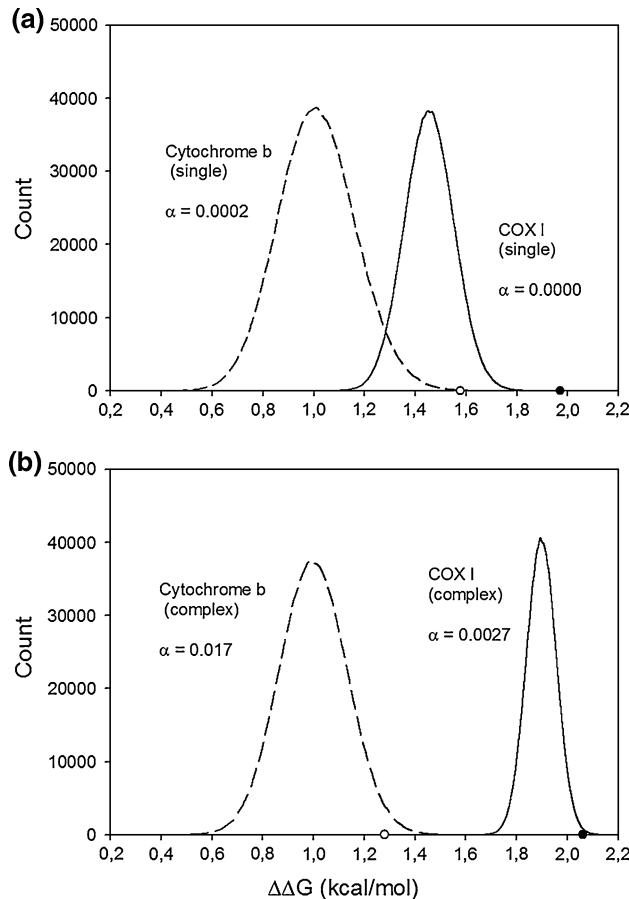


Fig. 6 Destabilizing effect of random mutations in cytochrome b and COX I. The mean free energy changes upon mutation of buried residues to alanine ($\Delta\Delta G$) were computed and indicated as open (cytochrome b) and filled circles (COX I) on the abscissae axes. These mean values were compared with the distribution of mean $\Delta\Delta G$ values of the same number of residues randomly chosen from the whole protein (see text for details). Direct comparisons of the distributions of mean $\Delta\Delta G$ in cytochrome b and COX I as single polypeptides (a) or as part of their respective complexes (b), are shown

2009; Franzosa and Xia 2009; Lin et al. 2007), have pointed to residue exposure as an important structural determinant of molecular evolution. In spite of a growing literature supporting the view that solvent exposed residues evolve faster than buried residues, this principle could not be taken for granted when mtDNA-encoded proteins are under consideration, because mitochondrial proteins often evolve under different selective constraints to those of nDNA-encoded proteins. This point is well illustrated by the observation that, while nDNA-encoded residues in the interface of protein complexes are more conserved, their mtDNA-encoded counterparts evolve even faster than other mtDNA-encoded residues (Schmidt et al. 2001). Furthermore, mtDNA encodes for membrane-spanning proteins, which seem to evolve differently than other proteins do (Conant et al. 2007; Popot and Engelman 2000), raising doubts about the applicability of the evolutionary rules that

govern soluble proteins (Conant and Stadler 2009). However, our results show that, for mtDNA-encoded proteins, the surfaces evolve faster than their corresponding interiors (Fig. 4e, f), providing evidence of the generality of this evolutionary rule across genomes and protein types.

Although these two regions also evolve differently in mitochondrial proteins, these differences are not so similar across proteins as it seems to be the case for yeast globular proteins (Conant 2009; Conant and Stadler 2009). In this respect, the interior of cytochrome b, in sharp contrast to that of COX I, shows a remarkable tolerance to changes as indicated by several lines of evidence. First, we failed to observe departures from random expectations in the distribution of invariant residues between the surface and the interior of cytochrome b (Table 1). Second, the mean Shannon's entropy for cytochrome b buried residues is not significantly lower than the mean value for a random set of residues (Fig. 2). Third, while unconstrained residues seem to be depleted from the COX I interior, they are randomly distributed between interior and surface of cytochrome b (Table 2). Fourth, although amino acid changes taking place in the interior of cytochrome b are indeed more constrained than those happening on its surface, the fact to be emphasized here is that the nonsynonymous substitution rates for the interior of cytochrome b were significantly higher than those computed for the surface of COX I (Fig. 5c). Finally, it should be noted that the thermodynamic stability changes of mutations affecting buried residues in cytochrome b are comparable to, or even lower than, those of exposed residues in COX I (Fig. 6).

While there is a consensus in the literature that exposed sites evolve faster than buried sites, the debate on whether the fraction of buried sites and d_N should correlate positively remains open. In other words, there is a wide agreement that solvent accessibility has a strong effect on the conservation of individual residues, but whether this behaviour scales to the level of whole proteins, is openly discussed. Since buried sites are more conserved, we may expect that proteins with a larger fraction of buried sites should evolve slower. On the other hand, it has been argued that although buried residues are generally more conserved, increasing the fraction of buried residues leads to an overall increase in the evolutionary rate of all residues in the protein (Bloom et al. 2006a). According to these authors, this is mainly because the higher number of buried residues yields a higher contact density that in turn contributes to an increase stability of the whole protein. Eventually, this additional stability allows a strong relaxation of constraint on the solved exposed sites. In few words, the reduction in the fraction of exposed residues is more than compensated for by the increased variability of exposed residues in proteins with high contact density. Although it may be the case for certain types of proteins, our results do not support

the generality of such arguments. Clearly, cytochrome b exhibits a much higher proportion of exposed residues than COX I, which is paralleled by a much higher rate of nonsynonymous substitutions. Therefore, the current analysis of mtDNA-encoded proteins lends support to the view of a positive correlation between the proportion of exposed residues and the rate of nonsynonymous substitutions (Lin et al. 2007).

It is widely acknowledged that mtDNA mutation rate can vary from gene to gene (Bielawski and Gold 1996). In fact, it has been suggested that the length of time genes remain in the single-stranded state during mtDNA replication (D_{ssH}), may be an important factor affecting the rate of mutation (Reyes et al. 1998). Since COX I and cytochrome b exhibit the shortest and longest D_{ssH} values, respectively, the superior rate of synonymous substitutions observed for COX I is a somewhat unexpected finding (Reyes et al. 1998). Nevertheless, this result could be explained by other mechanisms such as codon bias or time spent in the single-strand state during transcription (Faith and Pollock 2003).

Residues in the interior of the folded mitochondrial proteins and residues exposed on their surfaces, must experience different selection pressures as suggested by the observation that while the rates of synonymous substitutions are the same (Fig. 4c, d), there are great differences in rates of nonsynonymous changes (Fig. 4e, f). Nevertheless, the relevant question is, what are the biological factors behind this differential pattern of selection? In this sense, a main contribution of the current study is that we are providing compelling and quantitative evidence of the role played by thermodynamic stability. This evidence strongly suggests that the thermodynamic stability effect of mutations may be a key factor driving the evolutionary dynamics of proteins. Thus, we have observed that $\Delta\Delta G$ for mutations affecting buried residues are, in general, more destabilizing than those affecting exposed sites, which is in line with the lower d_N values computed for the interior with respect to the surface, regardless the protein being considered. It is worth noting that, despite the fact that residues from the interior of cytochrome b evolve under stronger constraints than residues exposed on its surface, buried residues from cytochrome b and exposed residues from COX I seem to evolve under comparable selective constraints (Fig. 5c). Interestingly, this observation fits well with the finding reported herein that $\Delta\Delta G$ for mutations affecting cytochrome b buried residues are comparable in magnitude to those for mutations affecting residues from the COX I surface. In few words, even the more destabilizing changes in cytochrome b interior are comparable in importance to the less destabilizing changes in COX I, which may help to explain why the core of cytochrome b seems to be much more tolerant to changes than the interior of COX I.

It has been suggested that the stability effect of mutations may show a universal energy distribution (Tokuriki et al. 2007). Therefore, the differential evolvability exhibited by proteins is believed to reside in the absolute thermodynamic stability of the native structure (Bloom et al. 2006b). In other words, random mutations indiscriminately decrease the stability of proteins. However, more stable proteins can tolerate better this decrease in stability, which in turn allows them to evolve faster. Somehow, our results challenge this simple view. We have shown that the magnitude of the stability effect of mutations may strongly depend on the tertiary structure of the protein under consideration. Furthermore, although we do not have data on the absolute stability of cytochrome b and COX I, if we accept the contact density as a proxy for absolute thermodynamic stability, we would expect COX I to be more stable than cytochrome b. However, the last evolves faster than the former. Therefore, we suggest that the stability effect of mutations strongly depends on the native structure and may be a key determinant of protein evolvability.

Conclusions

Although there exists a clear relationship between solvent exposure and the destabilizing effect of mutations (for a given protein, changes in the interior are more perturbing than changes at the surface), the absolute magnitude of the stability effect of mutations strongly depends on the native structure being considered. We suggest that $\Delta\Delta G$ rather than solvent accessibility may be the key determinant of the differential evolutionary behaviour of cytochrome b and COX I in mammals.

Acknowledgments We thank Alicia Esteban del Valle and Miguel Ángel Medina for their helpful comments on the manuscript. The authors are also grateful to two anonymous referees who have helped to improve the original manuscript. We also gratefully acknowledge the support of Grant CGL2010-18124 from the Ministerio de Ciencia e Innovación, Spain.

References

- Albu M, Min XJ, Hickey D, Golding B (2008) Uncorrected nucleotide bias in mtDNA can mimic the effects of positive Darwinian selection. *Mol Biol Evol* 25:2521–2524
- Aledo JC (2004) Glutamine breakdown in rapidly dividing cells: waste or investment? *Bioessays* 26:778–785
- Aledo JC, Li Y, de Magalhães JP, Ruiz-Camacho M, Pérez-Claros JA (2011) Mitochondrially encoded methionine is inversely related to longevity in mammals. *Aging Cell* 10:198–207
- Bielawski JP, Gold JR (1996) Unequal synonymous substitution rates within and between two protein-coding mitochondrial genes. *Mol Biol Evol* 13:889–892

- Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006a) Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol* 23:1751–1761
- Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006b) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874
- Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat Protoc* 4:1–13
- Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 76:1967–1971
- Bustamante CD, Townsend JP, Hartl DL (2000) Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol* 17:301–308
- Conant GC (2009) Neutral evolution on mammalian protein surfaces. *Trends Genet* 25:377–381
- Conant GC, Stadler PF (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol* 26:1155–1161
- Conant GC, Wagner GP, Stadler PF (2007) Modeling amino acid substitution patterns in orthologous and paralogous genes. *Mol Phylogenet Evol* 42:298–307
- DiMauro S, Schon EA (2008) Mitochondrial disorders in the nervous system. *Annu Rev Neurosci* 31:91–123
- Eilers M, Shekar SC, Shieh T, Smith SO, Fleming PJ (2000) Internal packing of helical membrane proteins. *Proc Natl Acad Sci USA* 97:5796–5801
- Faith JJ, Pollock DD (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165:735–745
- Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* 26:2387–2395
- Gallardo ME, Moreno-Loshuertos R, López C, Casqueiro M, Silva J, Bonilla F et al (2006) m.6267G>A: a recurrent mutation in the human mitochondrial DNA that reduces cytochrome c oxidase activity and is associated with tumors. *Hum Mutat* 27:575–582
- Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320:369–387
- Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267: 275–276
- Lin Y, Hsu W, Hwang J, Li W (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol* 24:1005–1011
- Miller S, Janin J, Lesk AM, Chothia C (1987) Interior and surface of monomeric proteins. *J Mol Biol* 196:641–656
- Mirny L, Shakhnovich E (2001) Evolutionary conservation of the folding nucleus. *J Mol Biol* 308:123–129
- Navarro A, Boveris A (2007) The mitochondrial energy transduction system and the aging process. *Am J Physiol* 292:C670–C686
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426
- Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7:337–348
- Popot JL, Engelman D (2000) Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem* 69:881–922
- Reyes A, Gissi C, Pesole G, Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 15:957–966
- Scharfe C, Lu HH, Neuenburg JK, Allen EA, Li G, Klopstock T et al (2009) Mapping gene associations in human mitochondria using clinical disease phenotypes. *PLoS Comput Biol* 5(4):e1000374
- Schmidt TR, Wu W, Goodman M, Grossman LI (2001) Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Mol Biol Evol* 18:563–569
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* 33:W382–W388
- Thompson MJ, Goldstein RA (1996) Constructing amino acid residue substitution classes maximally indicative of local protein structure. *Proteins* 25:28–37
- Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19:596–604
- Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369:1318–1332
- Tsodikov OV, Record MT, Sergeev YV (2002) Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J Comput Chem* 23:600–609
- Warnecke T, Weber CC, Hurst LD (2009) Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Trans* 37:756–761
- Welch JJ, Bininda-Emonds ORP, Bromham L (2008) Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol Biol* 8:53
- Zhang J, He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22:1147–1155

**E.3 Publicación del capítulo 3
(Bajo revisión)**

Evolution of residues from the cytochrome c oxidase complex engaged in intermolecular contacts

Héctor Valverde¹, Manuel Ruiz-Camacho², Ian Morilla¹, Francisco Demetrio López², Juan Carlos Aledo^{1,3}

Keywords: Coevolution, COX, evolvability, mtDNA, natural selection, protein evolution, protein-protein interaction.

Abstract

Respiratory complexes are encoded by two genomes (mtDNA and nDNA). Although the importance of intergenomic coadaptation is acknowledged, the forces and constraints shaping such coevolution are largely unknown. Previous works using cytochrome c oxidase (COX) as a model enzyme, have led to the so-called “optimizing interaction” hypothesis. According to this view, mtDNA-encoded residues close to nDNA-encoded residues evolve faster than the rest of positions, favouring the optimization of protein-protein interfaces. Herein, using evolutionary data in combination with structural information of COX, we show that failing to discern the effects of interaction from other structural effects, can lead to deceptive conclusions such as the “optimizing hypothesis”. Once spurious factors have been accounted for, data analysis shows that mtDNA-encoded residues engaged in contacts are, in general, more constrained than their noncontact counterpart. Nevertheless, noncontact residues from the surface of COX 1 subunit are

¹Departamento de Biología Molecular y Bioquímica. Facultad de Ciencias. Universidad de Málaga, 29071 Málaga, Spain

²Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Málaga, 29071 Málaga, Spain

³Corresponding author: caledo@uma.es

E. Publicaciones

a remarkable exception, being subjected to an exceptionally high purifying selection. We also report that mtDNA-encoded residues involved in contacts with other mtDNA-encoded subunits, are more constrained than mtDNA-encoded residues interacting with nDNA-encoded polypeptides. This differential behaviour cannot be explained on the basis of thermodynamic stability, since interactions between mtDNA-encoded subunits contribute more weakly to the complex stability than those interactions between subunits encoded by different genomes. Finally, among nDNA-encoded subunits, contact residues are more conserved than noncontact residues, being COX 5A an exception.

Introduction

Although mitochondria are involved in many aspects of cell function, including proliferation (Aledo 2004), apoptosis (Suen et al. 2008) and aging (Aledo et al. 2010), their central role is related to energy transduction in oxidative phosphorylation (OXPHOS). The mitochondrial proteins responsible for the OXPHOS are encoded by two genomes. In mammals, the mitochondrial genome (mtDNA) encodes for 13 polypeptides that interact with a large number of nuclear-encoded (nDNA-encoded) polypeptides to form the functional complexes I, III, IV and V of the OXPHOS system. Given that each mtDNA gene product must interact with proteins encoded by the nuclear genome to carry out its functions, coevolution between mtDNA and nDNA leading to intergenomic coadaptation is expected.

As a consequence of their critical function, mutations altering the structure of these oligomeric complexes must face the close scrutiny of natural selection. In this way, natural selection will favour evolutionary coadaptation of interacting proteins, either to improve physiological functions (Schmidt et al. 2005), or just to maintain the fitness through compensatory changes after a slightly deleterious mutation has been fixed by genetic drift (Osada and Akashi 2011). Whatever the driving force may be, there are numerous examples of the biological importance of intergenomic coadaptation. In this sense, xenomitochondrial cybrid cells constructed using nDNA from one species and mtDNA from a close species were viable and had a functional OXPHOS, whereas more divergent species failed to produce functional OXPHOS complexes (Kenyon and Moraes 1997; Barrientos et al. 2000; McKenzie 2003). These results underline the importance of cytonuclear coevolution. Similar conclusions were derived from studies where repeated backcrossing of genetically isolated populations of the copepod *Tigriopus californicus*, allowed to place the maternally inherited mtDNA genome of one population with the paternal nDNA of another. These interpopulation hybrids exhibited a defective OXPHOS system (Edmands and Burton 1999). A third line of evidence suggesting that OXPHOS proteins have coevolved to function optimally, comes from epistatic studies where mutations in mtDNA genes that are pathogenic in humans have been observed in naturally occurring genomes from nonhuman mammals (de Magalhães 2005; Azevedo et al. 2009).

While the above studies emphasize the relevance of coevolution between interacting proteins, they do not provide much insight on the forces and constraints shaping such coevolution. In an attempt to explore the evolutionary dynamics of these protein-protein interactions, Schmidt and co-workers used cytochrome c oxidase (COX) as a model of OXPHOS holoenzyme. These authors analysed the rate of nonsynonymous substitutions within a set formed by mtDNA-encoded residues in physical proximity to nDNA-encoded amino acids, and compared it with that computed for the rest of mtDNA-encoded residues, which are not in contact with nDNA-encoded polypeptide chains. They concluded that mtDNA-encoded residues in close contact with amino acids being encoded by the nucleus, evolve faster than the rest of mtDNA-encoded residues (Schmidt et al. 2001). This result was interpreted as being due to many different amino acid replacements among the close contact residues being required to optimize this protein's interaction with other proteins. Actually, the authors referred to such a state as an "optimizing interaction". In contrast, when COX residues encoded by nDNA were segregated on the basis of proximity to mtDNA-encoded residues, and the rates of non-synonymous substitution were analysed, the conclusion reached was the opposite. That is, those nDNA-encoded residues in contact with mtDNA-encoded amino acids, evolve more slowly than the rest of nDNA-encoded residues (Schmidt et al. 2001).

These striking results have been often cited as an example of the differential forces driving mtDNA and nDNA evolution (Willett 2003; Das et al. 2004; Castellana et al. 2011). While the constrained evolution of nDNA-encoded interacting residues is in line with the prevailing view that protein evolution is generally conservative and constraining interactions are typical, the observation that mtDNA-encoded interacting residues evolve at much higher rates than non interacting amino acids, if confirmed as a bona fide observation, deserved a sound explanation. Herein, we have revisited the "optimizing interaction" hypothesis. Using a comprehensive number of mammalian taxa and extended statistic analyses (Fig. E.1), we have found that interacting mtDNA-encoded residues, alike interacting nDNA-encoded residues, are subjected to higher constraints than their corresponding non-interacting counterparts. We also provide the keys to understand why previous studies failed to reach similar conclusions. In addition, we show an intriguing difference in the evolutionary rate of mtDNA-encoded residues depending on whether they contact with nDNA-encoded subunits or with other mtDNA- encoded subunits.

Results and discussion

Discerning the effects of residue interaction on the evolutionary rate from other structural effects.

The optimizing interaction hypothesis arose from the observation that mtDNA-encoded residues from COX in close contact with nDNA-encoded amino acids, showed a higher

E. Publicaciones

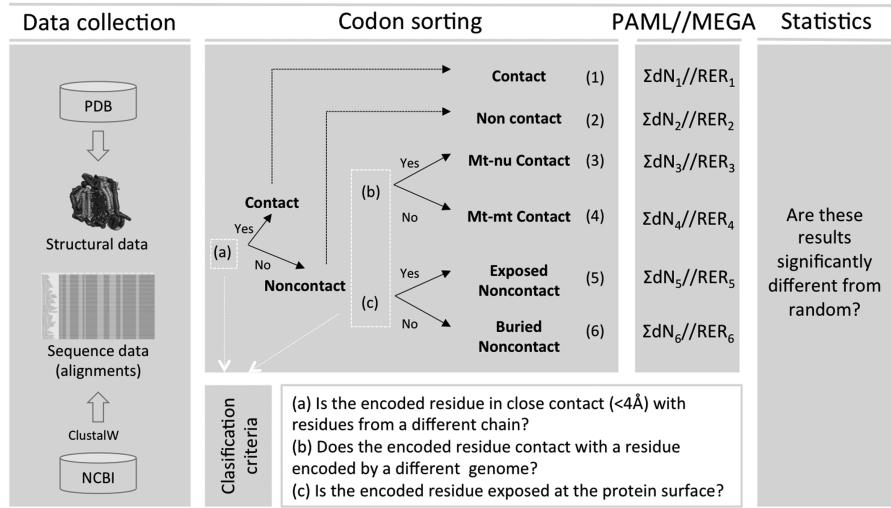


Figure E.1: Flowchart for the main methodological procedure adopted. Once sequence and structural data were collected, aligned codons were sorted into different subsets according to the criteria sketched in the figure (details are given in the text). Afterwards, pairwise nonsynonymous sequence divergences were calculated for each subset using PAML. The sum for all comparisons yields ΣdN_i , where the subscript i denotes the subset. In addition, the mean relative evolutionary rate (RER) for each subset was computed using MEGA5. Finally, to assess whether the values of these variables were significantly different between subsets, suitable statistical tests, which are described through the text, were carried out.

rate of nonsynonymous substitutions than the rest of mtDNA-encoded residues (Schmidt et al. 2001). Since these rates of nonsynonymous substitutions were originally calculated using orthologous sequences from only 26 mammalian species, we wanted to start by re-assessing this issue using a more comprehensive collection of sequences from 371 mammalian taxa. In this way, using our extensive alignment and the methodology described by Schmidt et al., the set formed by the aligned codons from the three mtDNA-encoded COX subunits (ABC), was split into two subsets: $ABC_{Mt-nu.Contact}$ and its complement, $(ABC_{Mt-nu.Contact})^c$. The former encompassed those triplets encoding for residues close to nDNA-encoded aminoacids in the holoenzyme, while the latter set contained all the codons from chain A, B and C, which encoded amino acids that are not in contact with nDNA-encoded residues (Fig. E.2). The calculated interaction ratio, $R_{ABC_{Mt-nu.Contact},(ABC_{Mt-nu.Contact})^c}$, was 1.805 ± 10^{-4} , significantly greater than 1. Although this result is in line with the data reported by Schmidt and coworkers, such an observation by itself is, in our opinion, insufficient to support the conclusion that contact residues are subjected to a positive selection, as suggested in previous works (Schmidt et al. 2001).

In this sense, a number of considerations need to be addressed before any conclusion can be reached. For instance, the set formed by mtDNA-encoded residues that are not in contact with nDNA-encoded amino acids, used as reference to compute the interaction ratio, represents a heterogeneous collection of residues. Thus, while the amino acids

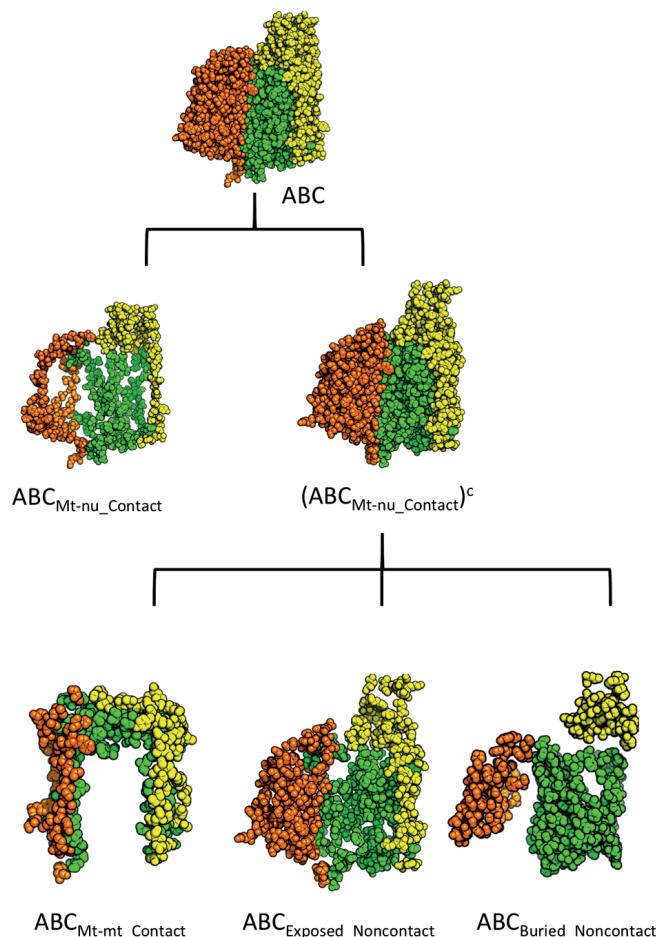


Figure E.2: Structural view of mitochondrially encoded COX residues. COX core, consisting of COX subunits 1 (chain A in green), 2 (chain B in yellow) and 3 (chain C in orange) is shown on the top of the figure. The spatial distribution of those residues in close contact with nDNA-encoded subunits ($\text{ABC}_{\text{Mt-nu_Contact}}$) is also shown. The set formed by mtDNA-encoded residues that are not in contact with nDNA-encoded subunits, $(\text{ABC}_{\text{Mt-nu_Contact}})^c$, was partitioned into three disjoint subsets: $\text{ABC}_{\text{Mt-mt_Contact}}$, which is formed by those residues contacting only with other mtDNA-encoded residues; $\text{ABC}_{\text{Exposed_Noncontact}}$, encompassing residues accessible to the solvent that are not involved in intersubunit contacts; and $\text{ABC}_{\text{Buried_Noncontact}}$, which contains all those residues that being buried inside the protein are not available for interusubunit contacts. The spatial distributions of the residues belonging to each of these subsets are shown at the bottom of the figure.

E. Publicaciones

belonging to the $\text{ABC}_{\text{Mt}-\text{nu}, \text{Contact}}$ group are mainly located at the protein surface, a significant part of the noncontact residues are buried into the protein structure (Fig. E.2). This observation is relevant because we have recently reported that, for mtDNA-encoded proteins, alike proteins from other genetic origins, buried residues are most likely to remain conserved during evolution compared to their solvent exposed counterparts (Aledo et al. 2012). Therefore, if we want to address the effects of residue interaction on the evolutionary rate, and discern them from other structural effects such as the solvent exposure, the noncontact set used as reference should be restricted to avoid buried residues.

Although the above mentioned restriction is necessary, yet it is not sufficient to build a suitable reference set. Indeed, the resulting set of such a restriction still contains a group of residues that may bias the results and mislead the conclusions. We are referring to the collection formed by those amino acids implicated in protein-protein interactions involving only contacts between mtDNA-encoded residues (see $\text{ABC}_{\text{Mt}-\text{mt}, \text{Contact}}$ in Figure E.2). Since the evolvability of this category of residues has not been previously characterized, we excluded them from the reference set, which finally was formed only by those mtDNA-encoded residues exposed at the protein surface that are not involved in any sort of intersubunit contacts. This reference set is referred to as $\text{ABC}_{\text{Exposed-Noncontact}}$ (Fig. E.2).

Once the partition of the initial data set had been carried out as described above and illustrated in Figure E.2, we next computed diverse evolutionary variables to characterize the relative evolvability of the following three sets: $\text{ABC}_{\text{Mt}-\text{nu}, \text{Contact}}$, $\text{ABC}_{\text{Mt}-\text{mt}, \text{Contact}}$ and $\text{ABC}_{\text{Exposed-Noncontact}}$, all of them containing only codons for amino acids exposed at the protein surface. The results of such analyses are described next.

Differential behaviour between Mt-mt and Mt-nu interactions.

To assess the relative evolvability of each category of residue we used different approaches. One of these, consisted in calculating the so-called relative evolutionary rate (RER). Briefly, for each mtDNA-encoded subunit, the RER of each position was computed using MEGA 5 (Tamura et al. 2011), which implements a maximum likelihood method to model evolutionary rate differences among sites. These rates were scaled such that the average evolutionary rate across all sites is 1. In this way, sites showing a rate lower than 1 are evolving slower than average, while those with a rate higher than 1 are evolving faster than average. As it can be observed in Table E.1, among all exposed residues, those involved in interactions between different mtDNA-encoded chains were the most constrained, showing RER values well below the unit, regardless the mtDNA-encoded subunit being considered (Table E.1). This observation indicates a much stronger purifying selection among Mt-mt Contact residues than among Mt-nu Contact amino acids. To substantiate this conclusion, a different approach was followed. For each of the three subsets of exposed residues (Exposed Noncontact, Mt-nu Contact and Mt-mt Contact) as well as for the Buried Noncontact subset, the num-

ber of nonsynonymous substitutions per nonsynonymous site, dN , was calculated from pairwise comparisons. The sum of these nonsynonymous sequence divergences for all the pairwise comparisons was computed and denoted as ΣdN . In line with the results based on RER, the Mt-mt Contact set exhibited the lower ΣdN values among all the exposed residues (Table E.1). Since dN is informative of the combined effect of mutation and selection, we used ΣdN as a proxy of the evolvability of the corresponding subset of residues being analysed. However, ΣdN averages over all the analysed mammalian species and, therefore, it provides little information on whether the evolutionary dynamics described above, has a broad phylogenetic distribution and is present in most mammalian lineages. To address this issue, phylogenetic trees were used to apportion the nonsynonymous substitutions for the different residue subsets. As it can be deduced from Fig. E.3, all lineages without exception showed much higher dN when the analysed subset was mtDNA-encoded residues contacting with nDNA-encoded subunits, with respect to those contacting with other mtDNA-encoded polypeptides. In most cases, the difference in the dN values computed for these two groups of residues was of one order of magnitude. (See also Supplementary Material-1). All together, these results suggest that those mtDNA-encoded residues in contact with mtDNA-encoded residues from a different chain, are subjected to stronger constraints than those involved in interactions with nDNA-encoded subunits. In addition, this behaviour seems to have a broad phylogenetic distribution and is valid for most mammalian lineages (Fig. E.3). To the best of our knowledge, this is the first quantitative study supporting such a conclusion, which may be rationalized as follows. In mammals, the three mtDNA-encoded subunits form the catalytic core of the enzyme, while the 10 nDNA-encoded subunits act as a regulatory shield surrounding the core (Soto et al. 2012). Prokaryote forms of COX boil down to the catalytic core, which seems to have an ancient origin (Castresana et al. 1994). In contrast, eukaryotes possess nuclear genes for additional subunits, the number of which generally increases with the organismal complexity. Although, neither the origin nor the specific function of these additional non-catalytic subunits is completely understood, there is little doubt that they are significantly younger than the proteins conforming the catalytic core (Castresana et al. 1994; Das et al. 2004; Little et al. 2010). On the other hand, it is well known that young proteins tend to experience weaker purifying selection and evolve more quickly than old proteins (Alba and Castresana 2005; Vishnoi et al. 2010). Herein, we propose that what is true for proteins may also be true for interactions. In other words, it is reasonable to assume that within a given protein with a fixed antiquity, those residues involved in “old interactions” evolve more slowly than those other residues implicated in “young interactions”. Describing how selection pressure acts at the interfaces of protein-protein complexes is a fundamental issue with high interest for the structural prediction of macromolecular assemblies. Therefore, although it is out of the scope of the current research, in the future it would be interesting to assess whether the relationship between the strength of selection and the age of the interaction, described herein for mtDNA-encoded COX subunits, also applies to other proteins. To encourage this task, we have released the program that implement part of the workflow sketched in Fig. E.1. This program has been submitted into the public repository CPAN

E. Publicaciones

under an open source license, which allows any other researcher to modify the software in order to include other structural, evolutionary and statistical analyses that may help to gain insights into the molecular evolution of any particular protein complex of interest. Documentation and guides for users and developers are available at the program website <http://mecom.hval.es>.

Thermodynamic stability does not account for the differential behaviour between Mt-mt and Mt-nu interactions.

Once we had established that residues from the Mt-mt Contact group are much more constrained than those belonging to the Mt-nu Contact set, we wondered if such differential pattern might have arisen from a differential contribution of these two categories of contacts to the stability of the complex. To address this issue, we carried out an *in silico* alanine scanning mutagenesis. The results of such analysis are summarized in Table E.2. As expected, changes affecting noncontact residues were among the least destabilizing mutations. On the other hand, from a thermodynamic point of view, residues from the Mt-nu Contact class were much less tolerant to changes than any other kind of residue, including those from the Mt-mt Contact set, which exhibited an intermediate behaviour. Since mutations between residues from the Mt-mt Contact class tend to be less destabilizing than those affecting residues from the Mt-nu Contact group, the higher degree of conservation observed between Mt-mt Contact residues, does not seem to be based on thermodynamic stability, suggesting that functional aspects may be behind the high degree of conservation observed among Mt-mt contact residues.

The exposed noncontact residues from COX 1 are exceptionally conserved.

Regardless of the genetic origin of the contacted residue, interacting mtDNA-encoded residues seem to be more constrained than their non interacting counterparts, as suggested by an interaction ratio below the unit ($0.8 \pm 2 \times 10^{-4}$). However, when each individual chain was separately analysed, COX 1 showed a unique behaviour. Thus, while the interaction ratios for COX 2 and COX 3 were significantly lower than 1, COX 1 exhibited a value significantly higher than 1. The departure of COX 1 from the general trend, may be due to either an increased rate of nonsynonymous substitutions between codons within the Contact set, which would favour the “optimizing hypothesis”, or to a reduced rate of nonsynonymous substitutions between codons belonging to the Exposed Noncontact set. To address this issue, we next computed and plotted ΣdN for the Contact and Exposed Noncontact sets of each mtDNA-encoded chain (Fig. E.4). From this figure, it becomes evident that i) changes in COX 1 are much more constrained than in any other chain, and ii) this was particularly true within the exposed noncontact group, arguing against the “optimizing hypothesis” even in the case of COX 1.

Table E.1: Contact residues from mtDNA-encoded subunits evolve differentially depending on the genetic origin of the contacted residue.

	Σd_N			
	ABC	A	B	C
Exposed Noncontact	5954 \pm 10	2901 \pm 9	10605 \pm 104	8404 \pm 50
(Mt-mt) Contact	2508 \pm 7	939 \pm 5	3034 \pm 27	5820 \pm 83
(Mt-nu) Contact	7050 \pm 15	7603 \pm 40	7129 \pm 56	6450 \pm 44
Buried Noncontact	1968 \pm 3	1252 \pm 4	5586 \pm 70	1230 \pm 14

	<i>RER</i>			
	ABC	A	B	C
Exposed Noncontact	1.27 \pm 2.26	1.03 \pm 2.17	1.52 \pm 1.87	1.48 \pm 2.54
(Mt-mt) Contact	0.41 \pm 0.92	0.26 \pm 0.31	0.32 \pm 0.45	0.88 \pm 1.82
(Mt-nu) Contact	1.60 \pm 2.33	2.33 \pm 2.98	1.21 \pm 1.67	0.90 \pm 1.26
Buried Noncontact	0.41 \pm 0.62	0.39 \pm 0.53	0.78 \pm 1.00	0.16 \pm 0.22

In the upper table, the rates of nonsynonymous substitution per nonsynonymous site, d_N , were calculated from pairwise comparisons using the alignment data subsets indicated. The sum of these nonsynonymous sequence divergences for all the pairwise comparisons was computed and denoted as Σd_N . Data are expressed as mean \pm standard deviation. In the lower table, a different approach was used to assess the relative evolvability of the different residue subsets. The relative evolutionary rate of each residue was computed using MEGA 5. These rates were scaled such that the average evolutionary rate across all sites is 1. This means that sites showing a rate lower than 1 are evolving slower than average, and those with a rate higher than 1 are evolving faster than average. A short movie showing multiple views of the protein surface formed by residues engaged in Mt-nu (magenta) and Mt-mt contacts (blue), as well as the surface of those residues that are not involved at all in intersubunit contacts (green), can be seen in Supplementary Material-5.

E. Publicaciones

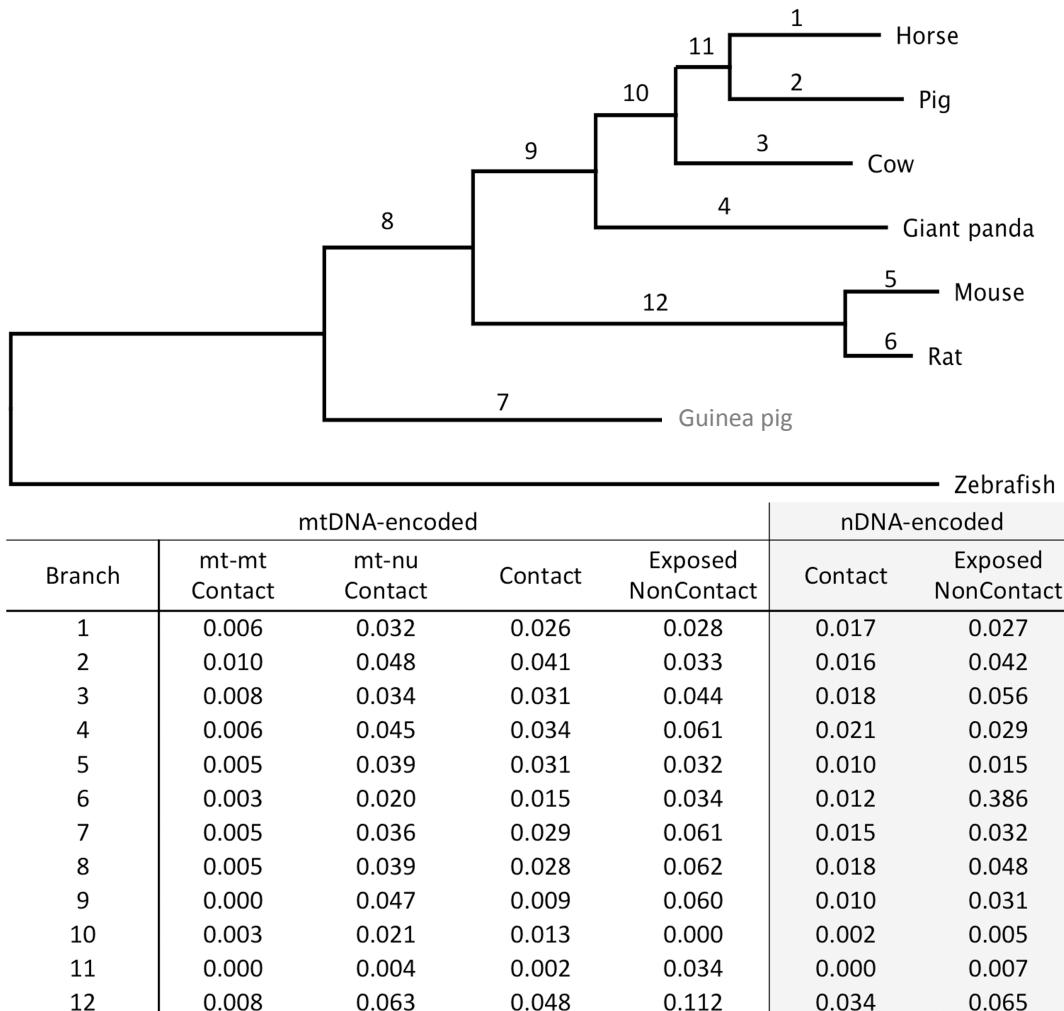


Figure E.3: Patterns of nonsynonymous substitutions in COX subunits across different mammalian lineages. The phylogenetic tree of seven mammalian species (*Bos taurus*, *Sus scrofa*, *Equus caballus*, *Ailuropoda melanoleuca*, *Mus musculus*, *Rattus norvegicus* and *Cavia porcellus*) for which the gene sequences of the 13 COX subunits were available, was reconstructed using *Arbacia lixula* as an outgroup. The number of nonsynonymous substitutions per nonsynonymous site for each branch is shown for the indicated residue subset.

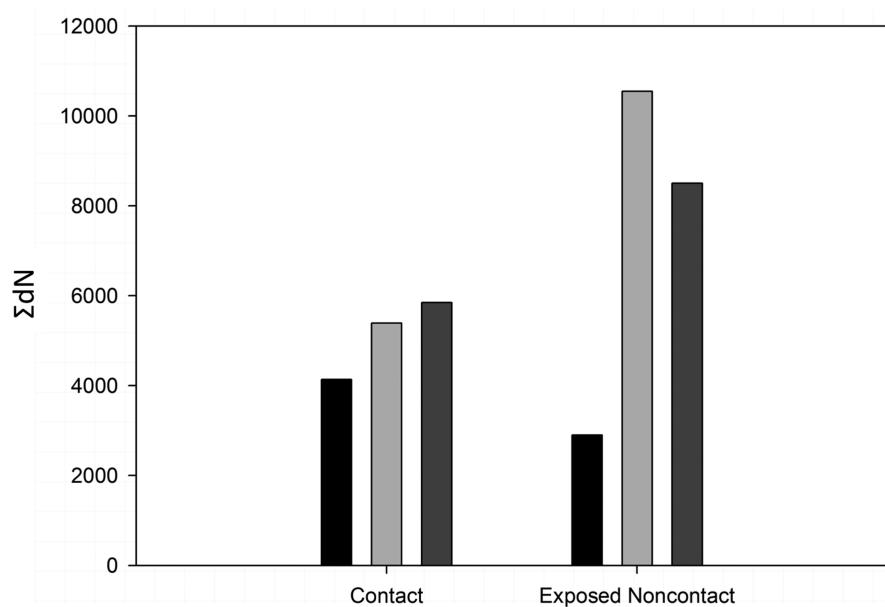


Figure E.4: Exposed noncontact residues from COX 1 are conspicuously conserved. Exposed noncontact residues from COX 1 are conspicuously conserved. The Contact and Exposed Noncontact sets from each mtDNA-encoded chain were used to compute the corresponding ΣdN values. Black, light grey and dark grey bars represent COX 1, 2 and 3, respectively. From this figure, it is evident that exposed noncontact residues from COX 1 exhibit little tendency to mutate.

E. Publicaciones

Table E.2: Thermodynamic stability changes.

	$\Delta\Delta G$ (kJ/mol)			
	ABC	A	B	C
Exposed Noncontact	1.41 ± 1.51 (†,§§)	1.64 ± 1.69 (†)	1.20 ± 1.09 (†,§§)	1.16 ± 1.38 (§)
(Mt-mt) Contact	1.78 ± 1.97 (*)	1.98 ± 1.65	1.92 ± 2.61	1.06 ± 1.52 (§)
(Mt-nu) Contact	1.92 ± 1.60 (**)	2.03 ± 1.55 (*)	1.99 ± 1.47 (**)	1.69 ± 1.73 (*,†)

The number of exposed noncontact residues from COX 1-3 were 144, 56 and 92, respectively. The number of mt-mt contact residues from COX 1-3 were 95, 54 and 37, respectively. The number of mt-nu contact residues from chain COX 1-3 were 120, 74 and 82, respectively. Data are expressed as mean ± standard error.

* Significantly different from Exposed Noncontact, (*) $p < 0.05$, (**) $p < 0.0005$.

† Significantly different from Mt-mt Contact, (†) $p < 0.05$, (††) $p < 0.0005$.

§ Significantly different from Mt-nu Contact, (§) $p < 0.05$, (§§) $p < 0.0005$.

From the comparison of ΣdN between subunits, it becomes clear-cut that exposed noncontact residues from COX 1 are particularly constrained (Fig. E.4). However, if we compare the inertia to changes of this subset of residues with respect to the rest of COX 1 residues, we may wonder whether exposed noncontact residues from COX 1 are significantly more conserved than the rest of COX 1 residues. To address this question while avoiding assumptions about the underlying distributions, we resorted to a bootstrap approach. Briefly, for each mtDNA-encoded chain, the codons from the multiple sequence alignment were randomly sorted to form a subset of the same size than the original Exposed Noncontact subset of the corresponding chain. Afterwards, ΣdN was computed using this random subset. For each chain, the random resampling was performed 10^4 times to build up empirical distributions, which were used to contrast the ΣdN values computed in the real Exposed Noncontact subsets. As it can be deduced from Fig. E.5, those residues belonging to the Exposed Noncontact groups from COX 2 and COX 3 are among the most variable residues (p-values 0.003 and 0.045, respectively). In contrast, exposed noncontact residues from COX 1 were among the most conserved residues (Fig. E.5).

In an attempt to get further insight into the particular forces imposing such exceptionally high degree of conservation between exposed noncontact COX 1 residues, we assessed the effect of in silico alanine scanning mutagenesis on the stability of the complex. Although the mean $\Delta\Delta G$ for exposed noncontact COX 1 residues (1.64 kJ/mol)

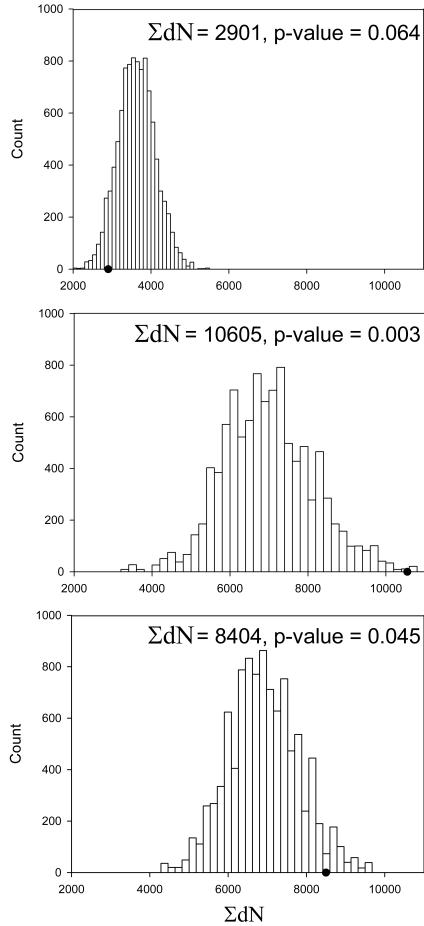


Figure E.5: The behaviour of exposed noncontact residues from COX 1 diverges from that exhibited for their counterparts in COX 2 and COX 3. For each mtDNA-encoded chain the codons from the multiple sequence alignment were randomly sorted to form a subset of the same size than the original Exposed Noncontact subset of the corresponding chain. Afterwards, ΣdN was computed using this random subset. For each chain, the random resampling was performed 10^4 times to build up empirical distributions, which were used to contrast the ΣdN values computed in the real Exposed Noncontact subsets, which are indicated as filled circles on the abscissa axis.

E. Publicaciones

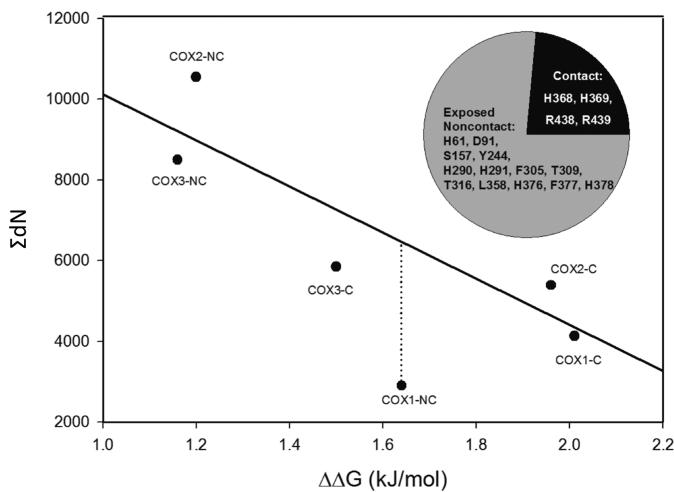


Figure E.6: Exposed noncontact residues from COX 1 are much more conserved than expected from their destabilizing effect on the holoenzyme. The exposed residues from each mtDNA-encoded proteins were split into two groups: contact (C) and noncontact (NC), according to the criteria given in the text. Afterwards, the ΣdN and mean $\Delta\Delta G$ were computed for each of these subsets. These two variables showed a significant negative correlation ($n = 6$, Pearson $r = -0.774$, p -value = 0.035) that was improved when data corresponding to COX 1 Exposed Noncontact were excluded from the analysis ($n=5$, Pearson $r = -0.890$, p -value = 0.021). The inset shows the proportion of residues that have been described as functionally relevant present among the Exposed Noncontact (grey), in comparison with the proportion of these residues that belong to the Contact group (black).

was significantly higher than those values from COX 2 (1.20 kJ/mol) and COX 3 (1.16 kJ/mol), p -values 0.015 and 0.008, respectively, the contribution of COX 1 exposed noncontact residues to the whole thermodynamic stability of the holoenzyme can hardly be invoked as a reason for their high degree of conservation, as it is evident from Fig. E.6. For each mtDNA-encoded protein, the Contact and Exposed Noncontact subsets were employed to compute their ΣdN values, which were plotted against their corresponding $\Delta\Delta G$ mean values (Fig. E.6). We found a significant (p -value = 0.035) negative correlation between these two variables, indicating that those substitutions that tend to be more destabilizing are also more constrained, which is in line with our previous observations that thermodynamic stability plays a relevant role in the evolvability of mtDNA-encoded proteins (Aledo et al. 2012). However, the COX 1 Exposed Noncontact group showed up as an outlier. In this sense, for a $\Delta\Delta G$ of 1.64 kJ/mol (the mean value computed for COX 1 exposed noncontact residues) the expected ΣdN should be higher than twice the observed ΣdN (Fig 6). In other words, whatever the constraining forces may be, they seem to be unrelated to structural stability.

Thus, we next explored other alternatives. In this sense, it is widely acknowledged that functional regions of proteins exhibit higher inertia to nonsynonymous changes than those other regions (Glaser et al. 2003). Hence, if there were a higher proportion

of functionally important residues in the exposed noncontact region of COX 1 than in its contact counterpart, this would help to explain the extraordinarily low ΣdN computed for the set Exposed Noncontact from COX 1. To test this possibility, we focused on those COX 1 residues that have been described to fulfil important functions such as the binding of heme, Cu and Mg, proton pumping and electron transfer (Michel et al. 1998; Yoshikawa 1998). Figure E.6 inset shows that the proportion of these residues that belong to the Exposed Noncontact set override the proportion of those that pertain to the Contact group. Although this observation is qualitatively in line with the very low ΣdN observed for the COX 1 Exposed Noncontact ensemble, it should be noted that the catalogue of COX 1 functional residues we have used, is very limited in size and probably it is far from the complete inventory, which hampers obtaining quantitative statistical support. Nevertheless, the exceptionally high degree of conservation observed between exposed noncontact COX 1 residues, likely captures important constraints that apply to biological functionality yet to be unravelled.

In this regard, we wondered whether this set of COX 1 exposed noncontact residues might be involved in functions related to the formation of mitochondrial supercomplexes. Indeed, during the past decade, significant experimental evidence supports the organization of the mitochondrial respiratory chain into higher order structures known as respirasome (Schägger and Pfeiffer 2000; Acín-Pérez et al. 2008; Winge 2012). Along the same lines, two quasi-simultaneous recent studies have reported the 3D structure of mitochondrial supercomplex I₁III₂IV₁, determined by electron cryo-microscopy at 19–22 Å resolution (Althoff et al. 2011; Dudkina et al. 2011). Fitting of X-ray structures of single complexes I, III₂ and IV with high fidelity, unravels only a few sites where neighbouring complexes come close enough for ion bridges or hydrogen bonds. Three of such sites of potentially strong protein-protein interaction were found between complexes III and IV. However, none of these sites were on COX 1 (Althoff et al. 2011; Dudkina et al. 2011). Nevertheless, beside the supercomplex I₁III₂IV₁, other forms of supercomplexes such as I₁III₂IV₂, I₁III₂IV₃ and I₁III₂IV₄, have been described in bovine heart mitochondria (Schägger and Pfeiffer 2001). Since the three dimensional structures of these less abundant forms of supercomplexes are unknown, the involvement of residues from COX 1 in the assembly/stability of respirasomes, although unlikely, cannot be completely ruled out. In any event, the high degree of conservation between the exposed noncontact residues from COX 1 herein described, is an unexpected and intriguing observation that we are currently unable to explain. Nevertheless, we have noted that when this particular behaviour of COX 1 noncontact residues is not properly accounted, it leads to high interaction ratios that may be misinterpreted as an optimizing high evolvability of Mt-nu contact residues.

E. Publicaciones

Contact residues from nDNA-encoded proteins tend to be more constrained than their noncontact counterparts.

A number of works have previously addressed the evolution of nDNA-encoded COX subunits (Schmidt et al. 2001; Das et al. 2004; Uddin et al. 2008; Osada and Akashi 2011). However, some of the conclusions drawn from these studies are in apparent conflict. For instance, Schmidt and coworkers concluded that while contact nDNA-encoded residues are subjected to a strong purifying selection (Schmidt et al. 2001), Osada and Akashi reached the conclusion that nDNA-encoded residues exhibit signs of positive selection (Osada and Akashi 2011). These conflicting results may be attributable to methodological differences. Thus, while the former authors assembled and analysed a data set formed by contact and noncontact residues without distinguishing between different COX subunits, the latter authors looked for individual sites under positive selection using the methodology described somewhere else (Zhang 2005). Herein, we have readdressed the issue employing a meso approach, consisting in analysing each individual nDNA-encoded chain.

As a first approach, we computed the interaction ratio for each subunit. Contact residues tend to be more conserved than their noncontact counterpart, as suggested by $R_{1,2}$ values well below the unit, with the exception of COX 5A (Fig. E.7A). This subunit is much more conserved than any of the other nuclear-encoded subunits (Fig. E.7B). The high level of conservation, which extends beyond the contact residues, suggests that subunit 5A has been subjected to an exceptionally strong purifying selection. Supporting this implication of functional constraint is experimental evidence for ligand interaction that involves subunit COX 5A as a receptor of 3,5-diiodothyronine, able to abolish the allosteric inhibition of respiration by ATP (Arnold et al. 1998).

In general lines, the results summarized in Fig. E.7A are consistent with those obtained by other authors using the pooled sequences (Schmidt et al. 2001). In analysing each individual nDNA-encoded polypeptide, we aimed to detect those subunits that might exhibit a particular evolutionary behaviour. However, such goal faced methodological challenges. Probably, the most important methodological limitation was the impossibility, because of the small size of the nDNA-encoded COX subunits, of obtaining a statistically relevant sample of residues. Therefore, in the absence of a large enough sample of residues, the herein computed $R_{1,2}$ values can only give a qualitative indication of the trend, but they lack of statistical significance. To overcome such inconvenience, we resorted to nonparametric analyses.

For a given subunit, each residue was labeled as 'variant' or 'invariant', depending on its conservation throughout a multisequence alignment (Supplementary Material-2). On the other hand, according to the criterion already described, the same residue was classified as 'contact' or 'exposed noncontact'. In this way, the proportion of variant exposed noncontact residues (p_0) and that of variant contact residues (p_1) were computed, as well as its differences, $d = p_0 - p_1$. According to a null hypothesis where the variability of a residue is independent of its role in intersubunit interaction, these

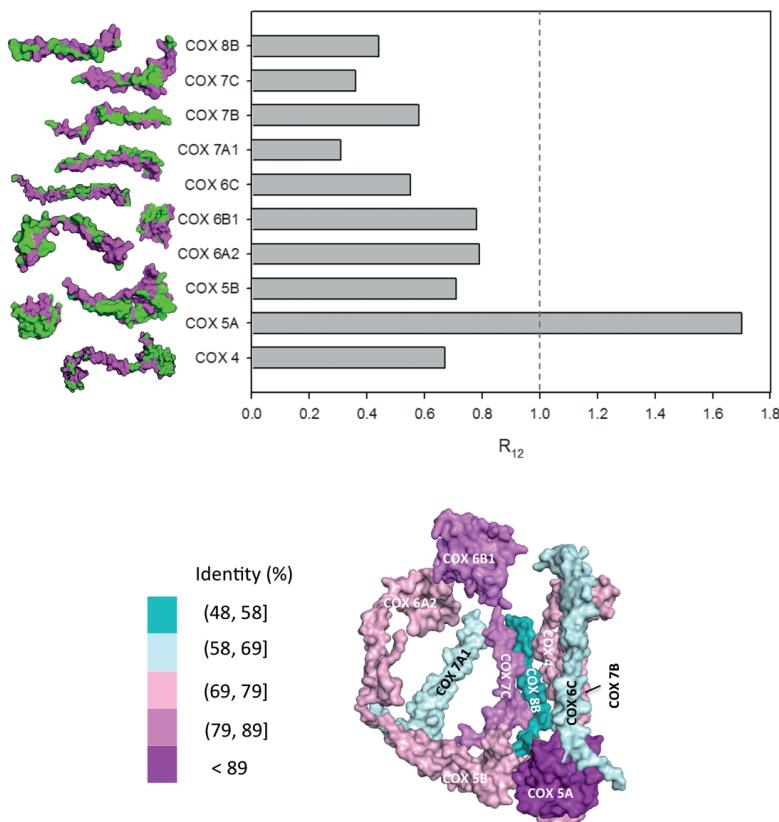


Figure E.7: COX 5A, a highly conserved protein, is the only nDNA-encoded subunit with an interaction ratio $R_{1,2}$ higher than the unit. (A) For each nDNA-encoded COX subunit, the ΣdN value for the Contact and Noncontact sets were computed to calculate the interaction ratio $R_{1,2}$. As it can be deduced from this figure, contact residues (magenta) tend to be more conserved than their noncontact counterpart (green), with the only exception of COX 5A. (B) To gain an indication of the relative rates of change among nuclearly encoded subunits, an alignment data set was constructed using only those species for which all the nDNA-encoded subunits were represented (*Ailuropoda melanoleuca*, *Bos taurus*, *Cavia porcellus*, *Equus caballus*, *Mus musculus*, *Rattus norvegicus* and *Sus scrofa*). Using this data set, the percentage of identity was computed for each subunit. Again COX 5A stood up, in this case as an exceptionally conserved protein.

E. Publicaciones

differences should fit to a symmetric distribution with zero median. To test such a hypothesis, a Wilcoxon signed-rank test was carried out, resulting in the rejection of the null hypothesis ($p < 0.007$). Thus, the higher proportion of variable residues among the exposed noncontact positions with respect to the contact positions (see Table from Supplementary Material-3) are unlikely to be due to chance.

Concluding remarks

Interactions of mtDNA- and nDNA-encoded proteins provide unique opportunities to study the evolution of protein-protein interactions and the effects of these interactions on the evolution of their respective genomes. To this respect, mammalian COX complex is well suited for studying the evolution of within- and between-genome interactions because a functional complex requires three mtDNA-encoded and ten nDNA-encoded subunits, accounting for a large number of interacting residues. Previous studies carried out with this complex, have suggested that mtDNA-encoded residues in close physical contact with nDNA- encoded amino acids may be subjected to positive selection to optimize the structural- functional interaction between subunits from different genetic origins. This conclusion was based on the higher rate of nonsynonymous substitutions observed among mtDNA-encoded residues in contact with nDNA-encoded residues, with respect to the rest of mtDNA-encoded amino acids. However, as we have shown herein, failing to discern the effect of interaction from other confounding factors such as the solvent exposure, can lead to misleading conclusions. Thus, when such corrections were made, data analysis showed that mtDNA- encoded residues engaged in contacts are, in general, more constrained than their noncontact counterparts. Nevertheless, noncontact residues from the surface of COX 1 subunit are a remarkable exception, being subjected to an exceptionally high purifying selection.

Besides providing compelling evidence against the so-called optimizing interaction hypothesis, our approach also allowed to make some interesting findings. For instance, those mtDNA-encoded residues in contact with mtDNA-encoded residues from a different chain, are subjected to stronger constrains than those involved in interactions with nDNA-encoded subunits. This observation cannot be explained on the basis of thermodynamic stability, because interactions between mtDNA-encoded subunits contribute more weakly to the complex stability than those interactions between subunits encoded by different genomes. Therefore, the higher conservation observed among mtDNA-encoded residues involved in within genome interactions, is likely due to functional rather than structural reasons.

Material and Methods

DNA Sequences

A collection of 371 mammalian mitochondrial genomes was obtained from the National Center for Biotechnology Information (NCBI) genome database (www.ncbi.nlm.nih.gov). For the analyses involving COX subunits encoded by nuclear genes, sequences from a variable number of species ranging from 14 to 30 (depending on sequence availability) were acquired from NCBI. A complete list of the mammalian taxa used and the accession number of the sequences is provided as Supplementary Material-4. Orthologous sequences were aligned by codons using ClustalW (the alignments can be retrieved from <http://mecom.hval.es>).

Codon sorting

Using the sequence from *Bos taurus* as reference and the crystal structure of bovine COX (Protein Data Bank, PDB, 2OCC), each codon position from the above described alignments, was sorted into different subsets according to the algorithm sketched in Figure E.1. Briefly, the data set corresponding to all the codons from the alignment of a given COX subunit (for instance, chain A, corresponding to COX 1, which is a mtDNA-encoded subunit), was initially divided into two subsets: Contact and Non-contact, depending on whether the encoded amino acid from chain A, is or not closer than 4\AA to a residue from any polypeptide other than chain A, respectively. Interacting positions were defined as being $< 4\text{\AA}$ apart because this is the upper limit for weak interactions (Martin et al. 1997). Afterwards, the Contact set was, in turn, split into two subsets: Intergenomic Contact (Mt-nu Contact, in the example) and Intragenomic Contact (Mt-mt Contact, in the example). The criterion to assign a given codon into the former subset was that the interacting residues should be encoded by different genomes, otherwise the codon is allocated into the latter subset. On the other hand, the Non-contact set was split up into two subsets: Exposed Noncontact and Buried Noncontact, on the basis of solvent accessible surface areas of the considered residue (Aledo et al. 2012). Computation of physical distances between residues and the sorting procedure was assisted by an *ad hoc* computer program that we have called MECOM (Molecular Evolution of protein COMplexes), which, together with the corresponding manual, can be downloaded from <http://mecom.hval.es>.

Nonsynonymous sequence divergences and statistics

Nonsynonymous sequence divergences were calculated from pairwise comparisons using the alignment data subsets described in the precedent section and employing the method of maximum likelihood implemented in PAML 4.6 (Yang and Nielsen 2000; Yang 2007). The sum of these nonsynonymous sequence divergences for all the pairwise comparisons

E. Publicaciones

was computed and denoted as ΣdN_i , where the subscript, i , refers to the subset used for the calculations (Fig. E.1).

Since we aimed to test for statistical evidence of the existence of differential evolutionary patterns among residues belonging to different subsets, we used the so-called interaction ratio (Schmidt et al. 2001), R , defined as the ratio between the ΣdN for the sets i being tested. For instance, if we want to compare residues involved in intersubunit contacts (subset 1) with the rest of residues, which are not implicated in such interactions (subset 2), we compute $R_{1,2} = \Sigma dN_1 / \Sigma dN_2$. In this way, a $R_{1,2}$ value indistinguishable from 1 should be interpreted as that both set of residues show similar nonsynonymous substitution rates. $R_{1,2} < 1$ would result from a reduced evolutionary rate among the residues belonging to subset 1, relative to those residues from subset 2. On the contrary, $R_{1,2} > 1$ points to a higher rate of nonsynonymous substitutions among residues from subset 1, with respect to those belonging to subset 2. To determine when $R_{i,j}$ differs from 1 because of a differential pattern of evolution and when it may be due to statistical uncertainty, a Z-test was performed to assess the significance of these deviations.

For mtDNA-encoded subunits (chains A, B and C corresponding to COX1, COX2 and COX3, respectively), which are the larger subunits from complex IV (514, 227 and 261 residues, respectively), it was possible to generate a reliable random distribution of ΣdN_i . To this end, the codons from multiple sequence alignments of each chain were randomly sorted into two subsets of the same sizes than the original subset i . Afterwards, ΣdN_i was computed as explained above. The random resampling was performed 10^4 times to build up an empirical distribution.

In order to calculate the number of nonsynonymous substitutions per nonsynonymous site on a lineage-by-lineage basis, we used a maximum likelihood method (F3x4 model) implemented in codeml from the PAML package (Yang 2007). To this end, the phylogenetic tree of seven mammalian species (*Bos taurus*, *Sus scrofa*, *Equus caballus*, *Ailuropoda melanoleuca*, *Mus musculus*, *Rattus norvegicus* and *Cavia porcellus*) for which the gene sequences of the 13 COX subunits were available, was reconstructed using *Arbacia lixula* as an outgroup. This tree and the alignments corresponding to the different residue categories were used to apportion the nonsynonymous substitutions among lineages.

Relative evolutionary rate

The codon alignment for each subunit was subjected to MEGA5 (Tamura et al. 2011) to calculate the relative evolutionary rate, RER. This software implements a maximum likelihood method and makes use of the Jones-Taylor-Thornton substitution model (Jones et al. 1992) to estimate the relative evolutionary rate at each single site. A discrete Gamma (+G) distribution was used to model evolutionary rate differences among sites. Five discrete Gamma categories were considered. The computed rates were scaled such that the average evolutionary rate across all sites is 1. This means that sites showing a

rate lower than 1 are evolving slower than average, and those with a rate higher than 1 are evolving faster than average.

Thermodynamic stability changes

The thermodynamic stability changes, $\Delta\Delta G$, of mutations were computed using the protein design tool FoldX version 3.0 (Guerois et al. 2002; Schymkowitz et al. 2005). FoldX uses a full atomic description of the structure of the protein, to provide a quantitative estimation of the importance of the interactions contributing to the stability of this protein. The 3D structure of COX was subjected to an optimization procedure using the repair function of FoldX. Afterwards, an alanine scan was carried out, the resulting $\Delta\Delta G$ were recorded and used to calculate the means for Exposed Noncontact, Mt-mt Contact and Mt-nu Contact residues.

Acknowledgments

This work was supported by grant CGL2010-18124 from the Ministerio de Ciencia e Innovación, Spain. We thank Alicia Esteban del Valle and Miguel Ángel Medina for their helpful comments on the manuscript.

References

- Acín-Pérez R, Fernández-Silva P, Peleato ML, Pérez-Martos A, Enríquez JA. 2008. Respiratory Active Mitochondrial Supercomplexes. *Mol Cell* 32:529-539.
- Alba MM, Catresana J. 2005. Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. *Mol Biol Evol* 22:598-606.
- Aledo JC. 2004. Glutamine breakdown in rapidly dividing cells: waste or investment? *Bioessays* 26:778-785.
- Aledo JC, Li Y, de Magalhães JP, Ruíz-Camacho M, Pérez-Claros JA. 2010. Mitochondrially encoded methionine is inversely related to longevity in mammals. *Aging Cell* 10:198- 207.
- Aledo JC, Valverde H, Ruíz-Camacho M. 2012. Thermodynamic Stability Explains the Differential Evolutionary Dynamics of Cytochrome b and COX I in Mammals. *J Mol Evol* 74:69-80.
- Althoff T, Mills DJ, Popot J-L, Kühlbrandt W. 2011. Arrangement of electron transport chain components in bovine mitochondrial supercomplex I1III2IV1. *EMBO J* 30:4652-4664.

E. Publicaciones

- Arnold S, Goglia F, Kadenbach B. 1998. 3,5-Diiodothyronine binds to subunit Va of cytochrome-c oxidase and abolishes the allosteric inhibition of respiration by ATP. *Eur J Biochem* 252:325-330.
- Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, Amorim A. 2009. Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. *BMC Genomics* 10:266.
- Barrientos A, Müller S, Dey R, Wienberg J, Morales CT (2000) Cytochrome c oxidase assembly in primates is sensitive to small evolutionary variations in amino acid sequence. *Mol. Biol. Evol.* 17, 1508-1519.
- Castellana S, Vicario S, Saccone C. 2011. Evolutionary Patterns of the Mitochondrial Genome in Metazoa: Exploring the Role of Mutation and Selection in Mitochondrial Protein-Coding Genes. *Genome Biol Evol* 3:1067-1079.
- Castresana J, Lübben M, Sarste M, Higgins DG. 1994. Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. *EMBO J* 13:2516-2525.
- Das J, Miller ST, Stern DL. 2004. Comparison of Diverse Protein Sequences of the Nuclear-Encoded Subunits of Cytochrome C Oxidase Suggests Conservation of Structure Underlies Evolving Functional Sites. *Mol Biol Evol* 21:1572-1582.
- Dudkina NV, Kudryashev M, Stahlberg H, Boekema EJ. 2011. Interaction of complexes I, III, and IV within the bovine respirasome by single particle cryoelectron tomography. *Proc. Natl Acad Sci USA* 108:15196-15200.
- Edmands S & Burton RS. 1999. Cytochrom c oxidase activity in interpopulation hybrids of a marine copepod: a test for nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution* 53:1972-1978.
- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. 2003. ConSurf: Identification of functional regions in proteins by surface-Mapping of phylogenetic information. *Bioinformatics* 19:163-164.
- Guerois R, Nielsen JE & Serrano L. 2002. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J Mol Biol* 320:369-387.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8:275-282.
- Kenyon L, Moraes CT. 1997. Expanding the functional human mitochondrial DNA database by the establishment of primate xenomitochondrial cybrids. *Proc Natl Acad Sci USA* 94:9131-9135.
- Little AG, Kocha KM, Lougheed SC, Moyes CD. 2010. Evolution of the nuclear-encoded cytochrome oxidase subunits in vertebrates. *Physiological Genomics* 42:76-

84.

de Magalhães JP. 2005. Human Disease-Associated Mitochondrial Mutations Fixed in Nonhuman Primates. *J Mol Evol* 61:491-497.

Martin PD, Michael GM, Box J, Esmon CT, Edwards BF. 1997. New insights into the regulation of the blood clotting cascade derived from the X-ray crystal structure of bovine meizothrombine des F1 in complex with PPACK. *Structure* 5:1681-1693.

McKenzie M. 2003. Functional Respiratory Chain Analyses in Murid Xenomitochondrial Cybrids Expose Coevolutionary Constraints of Cytochrome b and Nuclear Subunits of Complex III. *Mol Biol Evol* 20:1117-1124.

Michel H, Behr J, Harrenga A, Kannt A. 1998. Cytochrome c Oxidase: Structure and Spectroscopy. *Annu Rev Biophys Biomol Struct* 27:329-356.

Osada N, Akashi H. 2011. Mitochondrial-Nuclear Interactions and Accelerated Compensatory Evolution: Evidence from the Primate Cytochrome c Oxidase Complex. *Mol Biol Evol* 29:337-346.

Schägger H, Pfeiffer K. 2000. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *EMBO J* 19:1777-1783.

Schägger H, Pfeiffer K. 2001. The ratio of oxidative phosphorylation complexes I-V in bovine heart mitochondria and the composition of respiratory chain supercomplexes. *J Biol Chem* 276:37861-37867.

Schmidt TR, Wildman DE, Uddin M, Opazo JC, Goodman M & Grossman LI. 2005. Rapid electrostatic evolution at the binding site for cytochrome c on cytochrome c oxidase in anthropoid primates. *Proc Natl Acad Sci USA* 102:6379-6384.

Schmidt TR, Wu W, Goodman M, Grossman LI. 2001. Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Molecular Biology and Evolution* 18:563-569.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F & Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Research* 33:W382-W388.

Soto IC, Fontanesi F, Liu J & Barrientos A. 2012. Biogenesis and assembly of eukaryotic cytochrome c oxidase catalytic core. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1817:883-897.

Suen DF, Norris KL, Youle RJ. 2008. Mitochondrial dynamics and apoptosis. *Genes & Development* 22:1577-1590.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 28:2731-2739.

E. Publicaciones

- Uddin M, Opazo JC, Wildman DE, Sherwood CC, Hof PR, Goodman M, Grossman LI. 2008. Molecular evolution of the cytochrome c oxidase subunit 5A gene in primates. *BMC Evol Biol* 8:8.
- Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. 2010. Young proteins experience more variable selection pressures than old proteins. *Genome Res* 20:1574-1581.
- Willett CS. 2003. Evolution of Interacting Proteins in the Mitochondrial Electron Transport System in a Marine Copepod. *Mol Biol Evol* 21:443-453.
- Winge DR. 2012. Sealing the Mitochondrial Respirasome. *Mol Cell Biol* 32:2647-2652.
- Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24: 1586-1591.
- Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32-43.
- Yoshikawa S. 1998. Redox-Coupled Crystal Structural Changes in Bovine Heart Cytochrome c Oxidase. *Science* 280:1723-1729.
- Zhang J. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol* 22:2472-2479.

Apéndice F

Datos

F.1 Datos del capítulo 1

Tabla F.1: Genomas empleados en el trabajo recogido en el capítulo 1. En la tercera columna se indica el *accession number*¹ de la especie. La muestra se compone de los genomas mitocondriales de 173 especies de mamíferos, las cuales fueron clasificadas como longevas (L) o no longevas (S) dependiendo de si el *logMLSP* es mayor o menor a 2.46, valor que se corresponde con 228.4 meses. Adicionalmente, se muestra el logaritmo de la biomasa (*logBM*) en la 6^a columna.

no.	Especie	Accession number	Tipo	<i>logMLSP</i>	<i>logBM</i>
1	<i>Acinonyx jubatus</i>	NC_005212	S	2.3909	4.6733
2	<i>Ailuropoda melanoleuca</i>	NC_009492	S	2.6450	5.0107
3	<i>Ailurus fulgens</i>	NC_011124	S	2.3579	3.6990
4	<i>Ammotragus lervia</i>	NC_009510	S	2.4156	5.0212
5	<i>Arctocephalus pusillus</i>	NC_008417	S	2.5857	5.3010
6	<i>Artibeus jamaicensis</i>	NC_002009	L	2.3625	1.6484
7	<i>Balaena mysticetus</i>	NC_005268	S	3.4035	7.9624
8	<i>Balaenoptera acutorostrata</i>	NC_005271	S	2.7782	6.9191
9	<i>Balaenoptera borealis</i>	NC_006929	S	2.9484	7.3010
10	<i>Balaenoptera edeni</i>	NC_007938	S	2.9365	7.3010
11	<i>Balaenoptera musculus</i>	NC_001601	S	3.1206	8.2788
12	<i>Balaenoptera physalus</i>	NC_001321	S	3.1361	7.8451
13	<i>Berardius bairdii</i>	NC_005274	S	3.0035	7.0000
14	<i>Bos grunniens</i>	NC_006380	S	2.4991	5.7782
15	<i>Bos taurus</i>	NC_006853	S	2.3802	5.7875
16	<i>Bubalus bubalis</i>	NC_006295	S	2.6220	5.9633
17	<i>Camelus bactrianus</i>	NC_009628	S	2.6282	5.7653

¹El «*accession number*» es el código que identifica el genoma en la base de datos de NCBI (<http://www.ncbi.nlm.nih.gov/>).

F. Datos

no.	Especie	Accession number	Tipo	<i>logMLSP</i>	<i>logBM</i>
18	<i>Camelus dromedarius</i>	NC_009849	S	2.5325	5.7526
19	<i>Canis latrans</i>	NC_008093	S	2.4176	4.0959
20	<i>Canis lupus</i>	NC_008092	S	2.3930	4.6721
21	<i>Capra hircus</i>	NC_005044	S	2.3972	4.7076
22	<i>Cavia porcellus</i>	NC_000884	S	2.1584	2.8325
23	<i>Cebus albifrons</i>	NC_002763	S	2.6856	3.3424
24	<i>Ceratotherium simum</i>	NC_001808	S	2.7324	6.4370
25	<i>Cervus elaphus</i>	NC_007704	S	2.5775	5.2122
26	<i>Cervus nippon centralis</i>	NC_006993	S	2.4991	4.7243
27	<i>Cervus unicolor swinhoei</i>	NC_008414	S	2.5008	5.2553
28	<i>Chlorocebus aethiops</i>	NC_007009	S	2.5677	3.6128
29	<i>Choloepus didactylus</i>	NC_006924	S	2.6450	3.7574
30	<i>Colobus guereza</i>	NC_006901	S	2.6232	4.0086
31	<i>Cricetulus griseus</i>	NC_007936	L	1.8035	1.7782
32	<i>Cystophora cristata</i>	NC_008427	S	2.6232	5.5051
33	<i>Dactylopsila trivirgata</i>	NC_008134	S	2.0615	2.6160
34	<i>Dasyurus novemcinctus</i>	NC_001821	S	2.4275	3.6721
35	<i>Dasyurus hallucatus</i>	NC_007630	S	1.8500	2.6848
36	<i>Daubentonia madagascariensis</i>	NC_010299	S	2.4465	3.3979
37	<i>Didelphis virginiana</i>	NC_001610	S	1.8987	3.3959
38	<i>Distoechurus pennatus</i>	NC_008145	L	1.3802	1.7251
39	<i>Dugong dugon</i>	NC_003314	S	2.9425	5.5315
40	<i>Echinops telfairi</i>	NC_002631	L	2.3579	2.1038
41	<i>Echymipera rufescens australis</i>	NC_007632	S	1.5705	2.7896
42	<i>Elaphodus cephalophus</i>	NC_008749	S	2.4352	4.5250
43	<i>Elephantulus sp</i>	NC_004921	L	1.9868	1.7033
44	<i>Elephas maximus</i>	NC_005129	S	2.8954	6.5250
45	<i>Enhydra lutris</i>	NC_009692	S	2.5105	4.4232
46	<i>Equus asinus</i>	NC_001788	S	2.7513	5.2553
47	<i>Equus caballus</i>	NC_001640	S	2.8351	5.6021
48	<i>Erinaceus europaeus</i>	NC_002080	S	2.1474	2.8814
49	<i>Eschrichtius robustus</i>	NC_005270	S	2.9657	7.4548
50	<i>Eubalaena australis</i>	NC_006930	S	2.9243	7.6532
51	<i>Eubalaena japonica</i>	NC_006931	S	2.9243	7.8451
52	<i>Eulemur fulvus fulvus</i>	NC_012766	S	2.6294	3.4983
53	<i>Eulemur macaco macaco</i>	NC_012771	S	2.6379	3.3979
54	<i>Eulemur mongoz</i>	NC_010300	S	2.6379	3.3010
55	<i>Eumetopias jubatus</i>	NC_004030	S	2.5951	5.4914
56	<i>Felis catus</i>	NC_001700	S	2.5563	3.6675
57	<i>Galemys pyrenaicus</i>	NC_008156	L	1.7782	1.7709
58	<i>Glis glis</i>	NC_001892	L	2.0170	2.1903
59	<i>Gorilla gorilla</i>	NC_001645	S	2.8227	5.1847

F.1. Datos del capítulo 1

no.	Especie	Accession number	Tipo	$\log MLSP$	$\log BM$
60	<i>Gulo gulo</i>	NC_009685	S	2.3692	4.2304
61	<i>Halichoerus grypus</i>	NC_001602	S	2.7116	5.2455
62	<i>Helarctos malayanus</i>	NC_009968	S	2.6343	4.6628
63	<i>Hemiechinus auritus</i>	NC_005033	S	1.9600	2.5575
64	<i>Herpestes javanicus</i>	NC_006835	S	2.2993	2.7853
65	<i>Hippopotamus amphibius</i>	NC_000889	S	2.8659	6.4728
66	<i>Homo sapiens</i>	NC_012920	S	3.0792	4.7926
67	<i>Hydrurga leptonyx</i>	NC_008425	S	2.4942	5.5798
68	<i>Hylobates lar</i>	NC_002082	S	2.8274	3.6990
69	<i>Hyperoodon ampullatus</i>	NC_005273	S	2.6474	6.6335
70	<i>Inia geoffrensis</i>	NC_005276	S	2.5747	5.1335
71	<i>Isoodon macrourus</i>	NC_002746	S	1.9117	3.1906
72	<i>Jaculus jaculus</i>	NC_005314	L	1.9425	1.8129
73	<i>Kogia breviceps</i>	NC_005272	S	2.3096	5.6180
74	<i>Lama pacos</i>	NC_002504	S	2.4908	4.8129
75	<i>Lemur catta</i>	NC_004025	S	2.6509	3.4624
76	<i>Leptonychotes weddellii</i>	NC_008424	S	2.4771	5.6684
77	<i>Lepus europaeus</i>	NC_004028	S	2.1086	3.5911
78	<i>Lipotes vexillifer</i>	NC_007629	S	2.4594	4.9754
79	<i>Lobodon carcinophaga</i>	NC_008423	S	2.6702	5.3820
80	<i>Loxodonta africana</i>	NC_000934	S	2.8921	6.6021
81	<i>Lutra lutra</i>	NC_011358	S	2.3393	4.0000
82	<i>Macaca mulatta</i>	NC_005943	S	2.6812	3.8096
83	<i>Macaca sylvanus</i>	NC_002764	S	2.5431	4.0000
84	<i>Macropus robustus</i>	NC_001794	S	2.4216	4.4669
85	<i>Macroscelides proboscideus</i>	NC_004026	L	2.0187	1.6021
86	<i>Macrotis lagotis</i>	NC_006520	S	2.0615	3.1119
87	<i>Martes flavigula</i>	NC_012141	S	2.2833	3.6021
88	<i>Martes zibellina</i>	NC_011579	S	2.3440	3.0969
89	<i>Megaptera novaeangliae</i>	NC_006927	S	3.0569	7.4771
90	<i>Meles meles</i>	NC_011125	S	2.3487	4.0969
91	<i>Melursus ursinus</i>	NC_009970	S	2.6016	4.9759
92	<i>Metachirus nudicaudatus</i>	NC_006516	S	1.5563	2.5263
93	<i>Mirounga leonina</i>	NC_008422	S	2.4409	6.2041
94	<i>Mogera wogura</i>	NC_005035	L	1.5843	1.8319
95	<i>Monachus schauinslandi</i>	NC_008421	S	2.5563	5.3483
96	<i>Monodelphis domestica</i>	NC_006299	L	1.7868	2.0170
97	<i>Monodon monoceros</i>	NC_005279	S	2.7782	6.0314
98	<i>Muntiacus muntjak</i>	NC_004563	S	2.3533	4.2788
99	<i>Muntiacus reevesi</i>	NC_004069	S	2.4447	4.1761
100	<i>Mus musculus domesticus</i>	NC_006914	L	1.6812	1.2430
101	<i>Myrmecobius fasciatus</i>	NC_011949	S	2.1206	2.6021

F. Datos

no.	Especie	Accession number	Tipo	<i>logMLSP</i>	<i>logBM</i>
102	<i>Mystacina tuberculata</i>	NC_006925	L	1.9600	1.2553
103	<i>Nannospalax ehrenbergi</i>	NC_005315	L	2.2553	2.1553
104	<i>Nasalis larvatus</i>	NC_008216	S	2.4789	4.1004
105	<i>Neofelis nebulosa</i>	NC_008450	S	2.3758	4.3010
106	<i>Neophoca cinerea</i>	NC_008419	S	2.4612	5.2928
107	<i>Nycticebus coucang</i>	NC_002765	S	2.4908	3.0700
108	<i>Ochotona collaris</i>	NC_003033	L	1.8573	2.1139
109	<i>Ochotona princeps</i>	NC_005358	L	1.9243	2.0374
110	<i>Ornithorhynchus anatinus</i>	NC_000891	S	2.4333	2.8407
111	<i>Orycteropus afer</i>	NC_002078	S	2.5534	4.7005
112	<i>Oryctolagus cuniculus</i>	NC_001913	S	2.0334	3.3010
113	<i>Ovis aries</i>	NC_001941	S	2.4371	4.5340
114	<i>Pan paniscus</i>	NC_001644	S	2.8195	4.5623
115	<i>Pan troglodytes</i>	NC_001643	S	2.8530	4.6675
116	<i>Panthera pardus</i>	NC_010641	S	2.5153	4.7306
117	<i>Panthera tigris</i>	NC_010642	S	2.4991	5.1775
118	<i>Papio hamadryas</i>	NC_001992	S	2.6532	4.2385
119	<i>Pecari tajacu</i>	NC_012103	S	2.5775	4.3118
120	<i>Perameles gunnii</i>	NC_006521	S	1.8645	2.9227
121	<i>Petaurus breviceps</i>	NC_008135	L	2.3296	2.1038
122	<i>Phacochoerus africanus</i>	NC_008830	S	2.3993	4.9165
123	<i>Phascogale tapoatafa</i>	NC_006523	L	1.8500	2.1959
124	<i>Phascolarctos cinereus</i>	NC_008133	S	2.4236	3.6781
125	<i>Phoca hispida</i>	NC_008433	S	2.7419	4.9542
126	<i>Phoca largha</i>	NC_008430	S	2.7419	4.9590
127	<i>Phoca sibirica</i>	NC_008432	S	2.8274	5.0414
128	<i>Phoca vitulina</i>	NC_001325	S	2.7568	5.0212
129	<i>Phocoena phocoena</i>	NC_005280	S	2.3888	4.7202
130	<i>Physeter catodon</i>	NC_002503	S	2.9657	7.3617
131	<i>Pongo pygmaeus</i>	NC_001646	S	2.8500	4.7160
132	<i>Pontoporia blainvilliei</i>	NC_005277	S	2.2833	4.6075
133	<i>Potorous tridactylus</i>	NC_006524	S	2.2405	2.9894
134	<i>Presbytis melalophos</i>	NC_008217	S	2.3802	3.7993
135	<i>Procavia capensis</i>	NC_004919	S	2.2494	3.4771
136	<i>Procyon lotor</i>	NC_009126	S	2.4014	3.7324
137	<i>Pseudochirus peregrinus</i>	NC_006519	S	2.0835	2.9619
138	<i>Pteropus dasymallus</i>	NC_002612	S	2.4594	2.6675
139	<i>Pteropus scapulatus</i>	NC_002619	S	2.2778	2.6021
140	<i>Pygathrix nemaeus</i>	NC_008220	S	2.4942	3.9196
141	<i>Pygathrix roxellana</i>	NC_008218	S	2.5490	4.3617
142	<i>Rangifer tarandus</i>	NC_007703	S	2.4156	5.1335
143	<i>Rattus norvegicus</i>	NC_001665	S	1.7782	2.5551

F.1. Datos del capítulo 1

no.	Especie	Accession number	Tipo	$\log MLSP$	$\log BM$
144	<i>Rattus rattus</i>	NC_012374	L	1.7024	2.2788
145	<i>Rhinoceros unicornis</i>	NC_001779	S	2.7177	6.2304
146	<i>Rousettus aegyptiacus</i>	NC_007393	L	2.4390	2.0969
147	<i>Sciurus vulgaris</i>	NC_002369	S	2.4390	2.6684
148	<i>Semnopithecus entellus</i>	NC_008215	S	2.5416	4.0618
149	<i>Sminthopsis crassicaudata</i>	NC_007631	L	1.7782	1.2148
150	<i>Sorex unguiculatus</i>	NC_005435	L	1.2553	1.1492
151	<i>Spilogale putorius</i>	NC_010497	S	2.1004	2.7868
152	<i>Sus scrofa</i>	NC_000845	S	2.5105	4.9832
153	<i>Tachyglossus aculeatus</i>	NC_003321	S	2.7738	3.4354
154	<i>Talpa europaea</i>	NC_002391	L	1.9243	1.8865
155	<i>Tamandua tetradactyla</i>	NC_004032	S	2.3579	3.6537
156	<i>Tarsipes rostratus</i>	NC_006518	L	1.3802	1.0000
157	<i>Tarsius bancanus</i>	NC_002811	L	2.2914	1.9745
158	<i>Tarsius syrichta</i>	NC_012774	L	2.2833	2.0760
159	<i>Thryonomys swinderianus</i>	NC_002658	S	1.8116	3.7243
160	<i>Trachypithecus obscurus</i>	NC_006900	S	2.6094	3.8325
161	<i>Tremarctos ornatus</i>	NC_009969	S	2.6702	5.1263
162	<i>Trichechus manatus</i>	NC_010302	S	2.8274	5.6180
163	<i>Trichosurus vulpecula</i>	NC_003039	S	2.2806	3.4314
164	<i>Tupaia belangeri</i>	NC_002521	L	2.1245	2.3010
165	<i>Uncia uncia</i>	NC_010638	S	2.4055	4.6812
166	<i>Urotrichus talpoides</i>	NC_005034	L	1.6232	1.3096
167	<i>Ursus americanus</i>	NC_003426	S	2.6107	5.1761
168	<i>Ursus arctos</i>	NC_003427	S	2.6812	5.3222
169	<i>Ursus maritimus</i>	NC_003428	S	2.7207	5.5911
170	<i>Ursus thibetanus</i>	NC_009971	S	2.6725	4.9482
171	<i>Vombatus ursinus</i>	NC_003322	S	2.5563	4.4150
172	<i>Vulpes vulpes</i>	NC_008434	S	2.4076	3.7243
173	<i>Zaglossus bruijni</i>	NC_006364	S	2.6941	4.0128

F. Datos

Los análisis sobre las secuencias de DNA mitocondrial realizados en el capítulo 1, y cuyos resultados se muestran en la figura 1.5, se realizaron de igual forma sobre un conjunto de proteínas codificadas por nDNA que forman parte de la cadena respiratoria mitocondrial. En la tabla F.2 se muestran los identificadores de los genes que codifican dichas proteínas. Los resultados obtenidos tras estos análisis han sido incluidos en el apéndice C.

Tabla F.2: Catálogo de genes analizados codificados por nDNA. 38 del complejo I, 9 del complejo III y 14 del complejo IV. Estas secuencias fueron obtenidas de la base de datos KEGG (*Kyoto Encyclopedia of Genes and Genomes*, <http://www.genome.jp/kegg>).

no.	Gen	Id.
Complejo I		
1	NDUFS1	K03934
2	NDUFS2	K03935
3	NDUFS3	K03936
4	NDUFS4	K03937
5	NDUFS5	K03938
6	NDUFS6	K03939
7	NDUFS7	K03940
8	NDUFS8	K03941
9	NDUFV1	K03942
10	NDUFV2	K03943
11	NDUFV3	K03944
12	NDUFA1	K03945
13	NDUFA2	K03946
14	NDUFA3	K03947
15	NDUFA4	K03948
16	NDUFA5	K03949
17	NDUFA6	K03950
18	NDUFA7	K03951
19	NDUFA8	K03952
20	NDUFA9	K03953
21	NDUFA10	K03954
22	NDUFAB1	K03955

no.	Gen	Id.
23	NDUFA11	K03956
24	NDUFA12	K11352
25	NDUFA13	K11353
26	NDUFB1	K03957
27	NDUFB2	K03958
28	NDUFB3	K03959
29	NDUFB4	K03960
30	NDUFB5	K03961
31	NDUFB6	K03962
32	NDUFB7	K03963
33	NDUFB8	K03964
34	NDUFB9	K03965
35	NDUFB10	K03966
36	NDUFB11	K11351
37	NDUFC1	K03967
38	NDUFC2	K03968

Complejo III		
no.	Gen	Id.
1	UQCRRFS1	K00411
2	CYC1	K00413
3	QCR1	K00414
4	QCR2	K00415
5	QCR6	K00416
6	QCR7	K00417
7	QCR8	K00418
8	QCR9	K00419
9	QCR10	K00420

Complejo IV		
no.	Gen	Id.
1	COX10	K02257
2	COX4	K02263
3	COX5A	K02264
4	COX5B	K02265
5	COX6A	K02266
6	COX6B	K02267
7	COX6C	K02268
8	COX7	K02269

F.1. Datos del capítulo 1

no.	Gen	Id.
9	COX7B	K02269
10	COX7C	K02272
11	COX8	K02273
12	COX11	K02258
13	COX15	K02259
14	COX17	K0226

F. Datos

F.2 Datos del capítulo 2

Tabla F.3: Catálogo de las 311 especies de las que se recopilaron las secuencias genéticas (cDNA) de los genes citocromo b y COX 1. Tras llevar a cabo los alineamientos y la posterior depuración de los datos, la muestra final que se analizó fue de 231 especies. En la tercera columna se indica el *accession number* del genoma mitocondrial de la especie.

no.	Especie	Accession number	Accession number
1	<i>Acinonyx jubatus</i>	NC_005212	27 <i>Balaenoptera borealis</i>
2	<i>Ailuropoda melanoleuca</i>	NC_009492	28 <i>Balaenoptera brydei</i>
3	<i>Ailurus fulgens</i>	NC_011124	29 <i>Balaenoptera edeni</i>
4	<i>Ammotragus lervia</i>	NC_009510	30 <i>Balaenoptera musculus</i>
5	<i>Anomalurus sp</i>	NC_009056	31 <i>Balaenoptera physalus</i>
6	<i>Arctocephalus forsteri</i>	NC_004023	32 <i>Berardius bairdii</i>
7	<i>Arctocephalus pusillus</i>	NC_008417	33 <i>Bos grunniens</i>
8	<i>Artibeus jamaicensis</i>	NC_002009	34 <i>Bos indicus</i>
9	<i>Chrysocloris asiatica</i>	NC_004920	35 <i>Bos taurus</i>
10	<i>Eremitalpa granti</i>	NC_010304	36 <i>Bradypterus tridactylus</i>
11	<i>Ammotragus lervia</i>	NC_009510	37 <i>Bubalus bubalis</i>
12	<i>Bos indicus</i>	NC_005971	38 <i>Callorhinus ursinus</i>
13	<i>Bos taurus</i>	NC_006853	39 <i>Camelus bactrianus</i>
14	<i>Camelus bactrianus</i>	NC_009628	40 <i>Camelus dromedarius</i>
15	<i>Camelus dromedarius</i>	NC_009849	41 <i>Canis latrans</i>
16	<i>Capra hircus</i>	NC_005044	42 <i>Canis lupus</i>
17	<i>Cervus elaphus</i>	NC_007704	43 <i>Caperea marginata</i>
18	<i>Cervus unicolor</i>	NC_008414	44 <i>Capra hircus</i>
19	<i>Muntiacus muntjak</i>	NC_004563	45 <i>Arctocephalus forsteri</i>
20	<i>Muntiacus reevesi</i>	NC_004069	46 <i>Canis latrans</i>
21	<i>Ammotragus lervia</i>	NC_009510	47 <i>Canis lupus</i>
22	<i>Phacochoerus africanus</i>	NC_008830	48 <i>Enhydra lutris</i>
23	<i>Sus scrofa</i>	NC_000845	49 <i>Eumetopias jubatus</i>
24	<i>Balaena mysticetus</i>	NC_005268	50 <i>Halichoerus grypus</i>
25	<i>Balaenoptera acutorostrata</i>	NC_005271	51 <i>Leptonychotes weddellii</i>
26	<i>Balaenoptera bonaerensis</i>	NC_006926	52 <i>Lobodon carcinophaga</i>
			53 <i>Lutra lutra</i>
			54 <i>Martes melampus</i>
			55 <i>Martes zibellina</i>
			56 <i>Panthera pardus</i>
			57 <i>Panthera tigris</i>
			58 <i>Phoca fasciata</i>
			59 <i>Phoca groenlandica</i>
			60 <i>Phoca largha</i>
			61 <i>Phoca sibirica</i>
			62 <i>Phoca vitulina</i>
			63 <i>Phocarctos hookeri</i>
			64 <i>Ursus arctos</i>

F.2. Datos del capítulo 2

no.	Especie	Accession number	no.	Especie	Accession number
65	<i>Ursus maritimus</i>	NC_003428	99	<i>Colobus guereza</i>	NC_006901
66	<i>Zalophus californianus</i>	NC_008416	100	<i>Cricetulus griseus</i>	NC_007936
67	<i>Cavia porcellu</i>	NC_000884	101	<i>Crocidura russula</i>	NC_006893
68	<i>Cebus albifrons</i>	NC_002763	102	<i>Cystophora cristata</i>	NC_008427
69	<i>Ceratotherium simum</i>	NC_001808	103	<i>Dactylopsila trivirgata</i>	NC_008134
70	<i>Cervus elaphus</i>	NC_007704	104	<i>Dasyurus novemcinctus</i>	NC_001821
71	<i>Cervus nippon centralis</i>	NC_006993	105	<i>Dasyurus hallucatus</i>	NC_007630
72	<i>Cervus unicolor</i>	NC_008414	106	<i>Myrmecobius fasciatus</i>	NC_011949
73	<i>Balaenoptera acutorostrata</i>	NC_005271	107	<i>Phascogale tapoatafa</i>	NC_006523
74	<i>Balaenoptera bonaerensis</i>	NC_006926	108	<i>Sminthopsis douglasi</i>	NC_006517
75	<i>Balaenoptera brydei</i>	NC_006928	109	<i>Dasyurus hallucatus</i>	NC_007630
76	<i>Balaenoptera edeni</i>	NC_007938	110	<i>Daubentonia madagascariensis</i>	NC_010299
77	<i>Balaenoptera physalus</i>	NC_001321	111	<i>Dendrohyrax dorsalis</i>	NC_010301
78	<i>Berardius bairdii</i>	NC_005274	112	<i>Metachirus nudicaudatus</i>	NC_006516
79	<i>Eubalaena australis</i>	NC_006930	113	<i>Thylamys elegans</i>	NC_005825
80	<i>Eubalaena japonica</i>	NC_006931	114	<i>Didelphis virginiana</i>	NC_001610
81	<i>Hyperoodon ampullatus</i>	NC_005273	115	<i>Dactylopsila trivirgata</i>	NC_008134
82	<i>Kogia breviceps</i>	NC_005272	116	<i>Petaurus breviceps</i>	NC_008135
83	<i>Megaptera novaeangliae</i>	NC_006927	117	<i>Phalanger interpositus</i>	NC_008137
84	<i>Monodon monoceros</i>	NC_005279	118	<i>Phascolarctos cinereus</i>	NC_008133
85	<i>Phocoena phocoena</i>	NC_005280	119	<i>Pseudocheirus peregrinus</i>	NC_006519
86	<i>Physeter catodon</i>	NC_002503	120	<i>Tarsipes rostratus</i>	NC_006518
87	<i>Chalinolobus tuberculatus</i>	NC_002626	121	<i>Trichosurus vulpecula</i>	NC_003039
88	<i>Artibeus jamaicensis</i>	NC_002009	122	<i>Vombatus ursinus</i>	NC_003322
89	<i>Chalinolobus tuberculatus</i>	NC_002626	123	<i>Distoechurus pennatus</i>	NC_008145
90	<i>Mystacina tuberculata</i>	NC_006925	124	<i>Dugong dugon</i>	NC_003314
91	<i>Pipistrellus abramus</i>	NC_005436	125	<i>Echinops telfairi</i>	NC_002631
92	<i>Pteropus dasymallus</i>	NC_002612	126	<i>Echymipera rufescens australis</i>	NC_007632
93	<i>Pteropus scapulatus</i>	NC_002619	127	<i>Elaphodus cephalophus</i>	NC_008749
94	<i>Rhinolophus monoceros</i>	NC_005433	128	<i>Elephantulus sp</i>	NC_004921
95	<i>Rhinolophus pumilus</i>	NC_005434	129	<i>Elephas maximus</i>	NC_005129
96	<i>Chlorocebus aethiops</i>	NC_007009	130	<i>Enhydra lutris</i>	NC_009692
97	<i>Choloepus didactylus</i>	NC_006924	131	<i>Equus asinus</i>	NC_001788
98	<i>Chrysochloris asiatica</i>	NC_004920	132	<i>Equus caballus</i>	NC_001640

F. Datos

no.	Especie	Accession number	no.	Especie	Accession number
133	<i>Eremitalpa granti</i>	NC_010304	169	<i>Lepus europaeus</i>	NC_004028
134	<i>Erinaceus europaeus</i>	NC_002080	170	<i>Lipotes vexillifer</i>	NC_007629
135	<i>Hemiechinus auritus</i>	NC_005033	171	<i>Lobodon carcinophaga</i>	NC_008423
136	<i>Erinaceus europaeus</i>	NC_002080	172	<i>Loxodonta africana</i>	NC_000934
137	<i>Eschrichtius robustus</i>	NC_005270	173	<i>Lutra lutra</i>	NC_011358
138	<i>Eubalaena australis</i>	NC_006930	174	<i>Macaca mulatta</i>	NC_005943
139	<i>Eubalaena japonica</i>	NC_006931	175	<i>Macaca sylvanus</i>	NC_002764
140	<i>Eulemur fulvus fulvus</i>	NC_012766	176	<i>Macropus robustus</i>	NC_001794
141	<i>Eulemur fulvus mayottensis</i>	NC_012769	177	<i>Macroscelides proboscideus</i>	NC_004026
142	<i>Eulemur mongoz</i>	NC_010300	178	<i>Macrotis lagotis</i>	NC_006520
143	<i>Eumetopias jubatus</i>	NC_004030	179	<i>Mammuthus primigenius</i>	NC_007596
144	<i>Felis catus</i>	NC_001700	180	<i>Martes flavigula</i>	NC_012141
145	<i>Galemys pyrenaicus</i>	NC_008156	181	<i>Martes melampus</i>	NC_009678
146	<i>Glis glis</i>	NC_001892	182	<i>Martes zibellina</i>	NC_011579
147	<i>Gorilla gorilla</i>	NC_001645	183	<i>Megaptera novaeangliae</i>	NC_006927
148	<i>Gulo gulo</i>	NC_009685	184	<i>Meles meles anakuma</i>	NC_009677
149	<i>Halichoerus grypus</i>	NC_001602	185	<i>Meles meles</i>	NC_011125
150	<i>Helarctos malayanus</i>	NC_009968	186	<i>Melursus ursinus</i>	NC_009970
151	<i>Hemiechinus auritus</i>	NC_005033	187	<i>Metachirus nudicaudatus</i>	NC_006516
152	<i>Herpestes javanicus</i>	NC_006835	188	<i>Microtus kikuchii</i>	NC_003041
153	<i>Hippopotamus amphibius</i>	NC_000889	189	<i>Microtus rossiaemericus</i>	NC_008064
154	<i>Homo sapiens</i>	NC_012920	190	<i>Mirounga leonina</i>	NC_008422
155	<i>Hydrurga leptonyx</i>	NC_008425	191	<i>Mogera wogura</i>	NC_005035
156	<i>Hylobates lar</i>	NC_002082	192	<i>Monachus schauinslandi</i>	NC_008421
157	<i>Hyperoodon ampullatus</i>	NC_005273	193	<i>Monodelphis domestica</i>	NC_006299
158	<i>Dendrohyrax dorsalis</i>	NC_010301	194	<i>Monodon monoceros</i>	NC_005279
159	<i>Procavia capensis</i>	NC_004919	195	<i>Ornithorhynchus anatinus</i>	NC_000891
160	<i>Inia geoffrensis</i>	NC_005276	196	<i>Tachyglossus aculeatus</i>	NC_003321
161	<i>Isoodon macrourus</i>	NC_002746	197	<i>Muntiacus muntjak</i>	NC_004563
162	<i>Jaculus jaculus</i>	NC_005314	198	<i>Muntiacus reevesi</i>	NC_004069
163	<i>Kogia breviceps</i>	NC_005272	199	<i>Mus musculus domesticus</i>	NC_006914
164	<i>Ochotona collaris</i>	NC_003033			
165	<i>Ochotona princeps</i>	NC_005358			
166	<i>Lama pacos</i>	NC_002504			
167	<i>Lemur catta</i>	NC_004025			
168	<i>Leptonychotes weddellii</i>	NC_008424			

F.2. Datos del capítulo 2

no.	Especie	Accession number	no.	Especie	Accession number
200	<i>Myrmecobius fasciatus</i>	NC_011949	237	<i>Physeter catodon</i>	NC_002503
201	<i>Mystacina tuberculata</i>	NC_006925	238	<i>Bradypus tridactylus</i>	NC_006923
202	<i>Nannospalax ehrenbergi</i>	NC_005315	239	<i>Choloepus didactylus</i>	NC_006924
203	<i>Nasalis larvatus</i>	NC_008216	240	<i>Pipistrellus abramus</i>	NC_005436
204	<i>Neofelis nebulosa</i>	NC_008450	241	<i>Pongo abelii</i>	NC_002083
205	<i>Neophoca cinerea</i>	NC_008419	242	<i>Pongo pygmaeus</i>	NC_001646
206	<i>Nycticebus coucang</i>	NC_002765	243	<i>Pontoporia blainvilleyi</i>	NC_005277
207	<i>Ochotona collaris</i>	NC_003033	244	<i>Potorous tridactylus</i>	NC_006524
208	<i>Ochotona princeps</i>	NC_005358	245	<i>Presbytis melalophos</i>	NC_008217
209	<i>Ornithorhynchus anatinus</i>	NC_000891	246	<i>Colobus guereza</i>	NC_006901
210	<i>Orycteropus afer</i>	NC_002078	247	<i>Eulemur fulvus fulvus</i>	NC_012766
211	<i>Oryctolagus cuniculus</i>	NC_001913	248	<i>Eulemur fulvus mayottensis</i>	NC_012769
212	<i>Ovis aries</i>	NC_001941	249	<i>Macaca sylvanus</i>	NC_002764
213	<i>Pan paniscus</i>	NC_001644	250	<i>Nasalis larvatus</i>	NC_008216
214	<i>Pan troglodytes</i>	NC_001643	251	<i>Papio hamadryas</i>	NC_001992
215	<i>Panthera pardus</i>	NC_010641	252	<i>Pongo abelii</i>	NC_002083
216	<i>Panthera tigris</i>	NC_010642	253	<i>Pongo pygmaeus</i>	NC_001646
217	<i>Papio hamadryas</i>	NC_001992	254	<i>Procolobus badius</i>	NC_008219
218	<i>Isodon macrourus</i>	NC_002746	255	<i>Pygathrix roxellana</i>	NC_008218
219	<i>Macrotis lagotis</i>	NC_006520	256	<i>Tarsius bancanus</i>	NC_002811
220	<i>Pecari tajacu</i>	NC_012103	257	<i>Tarsius syrichta</i>	NC_012774
221	<i>Perameles gunnii</i>	NC_006521	258	<i>Elephas maximus</i>	NC_005129
222	<i>Ceratotherium simum</i>	NC_001808	259	<i>Mammuthus primigenius</i>	NC_007596
223	<i>Rhinoceros unicornis</i>	NC_001779	260	<i>Procavia capensis</i>	NC_004919
224	<i>Petaurus breviceps</i>	NC_008135	261	<i>Procolobus badius</i>	NC_008219
225	<i>Phacochoerus africanus</i>	NC_008830	262	<i>Procyon lotor</i>	NC_009126
226	<i>Phalanger interpositus</i>	NC_008137	263	<i>Pseudochirus peregrinus</i>	NC_006519
227	<i>Phascogale tapoatafa</i>	NC_006523	264	<i>Pteropus dasymallus</i>	NC_002612
228	<i>Phascolarctos cinereus</i>	NC_008133	265	<i>Pteropus scapulatus</i>	NC_002619
229	<i>Phoca fasciata</i>	NC_008428	266	<i>Pygathrix nemaeus</i>	NC_008220
230	<i>Phoca groenlandica</i>	NC_008429	267	<i>Pygathrix roxellana</i>	NC_008218
231	<i>Phoca hispida</i>	NC_008433	268	<i>Rangifer tarandus</i>	NC_007703
232	<i>Phoca largha</i>	NC_008430	269	<i>Rattus norvegicus</i>	NC_001665
233	<i>Phoca sibirica</i>	NC_008432	270	<i>Rattus rattus</i>	NC_012374
234	<i>Phoca vitulina</i>	NC_001325	271	<i>Rhinoceros unicornis</i>	NC_001779
235	<i>Phocarctos hookeri</i>	NC_008418			
236	<i>Phocoena phocoena</i>	NC_005280			

F. Datos

no.	Especie	Accession number	no.	Especie	Accession number
272	<i>Rhinolophus monoceros</i>	NC_005433	305	<i>Ursus arctos</i>	NC_003427
273	<i>Rhinolophus pumilus</i>	NC_005434	306	<i>Ursus maritimus</i>	NC_003428
274	<i>Microtus kikuchii</i>	NC_003041	307	<i>Ursus thibetanus</i>	NC_009971
275	<i>Microtus rossiaemeridionalis</i>	NC_008064	308	<i>Vombatus ursinus</i>	NC_003322
276	<i>Rattus norvegicus</i>	NC_001665	309	<i>Vulpes vulpes</i>	NC_008434
277	<i>Rattus rattus</i>	NC_012374	310	<i>Zaglossus bruijni</i>	NC_006364
278	<i>Sciurus vulgaris</i>	NC_002369	311	<i>Zalophus californianus</i>	NC_008416
279	<i>Thryonomys swinderianus</i>	NC_002658			
280	<i>Sciurus vulgaris</i>	NC_002369			
281	<i>Semnopithecus entellus</i>	NC_008215			
282	<i>Sminthopsis crassicaudata</i>	NC_007631			
283	<i>Sminthopsis douglasi</i>	NC_006517			
284	<i>Sorex unguiculatus</i>	NC_005435			
285	<i>Galemys pyrenaicus</i>	NC_008156			
286	<i>Urotrichus talpoides</i>	NC_005034			
287	<i>Spilogale putorius</i>	NC_010497			
288	<i>Sus scrofa</i>	NC_000845			
289	<i>Tachyglossus aculeatus</i>	NC_003321			
290	<i>Talpa europaea</i>	NC_002391			
291	<i>Tamandua tetradactyla</i>	NC_004032			
292	<i>Tarsipes rostratus</i>	NC_006518			
293	<i>Tarsius bancanus</i>	NC_002811			
294	<i>Tarsius syrichta</i>	NC_012774			
295	<i>Thryonomys swinderianus</i>	NC_002658			
296	<i>Thylamys elegans</i>	NC_005825			
297	<i>Trachypithecus obscurus</i>	NC_006900			
298	<i>Tremarctos ornatus</i>	NC_009969			
299	<i>Trichechus manatus</i>	NC_010302			
300	<i>Trichosurus vulpecula</i>	NC_003039			
301	<i>Tupaia belangeri</i>	NC_002521			
302	<i>Uncia uncia</i>	NC_010638			
303	<i>Urotrichus talpoides</i>	NC_005034			
304	<i>Ursus americanus</i>	NC_003426			

F.3. Datos del capítulo 3

F.3 Datos del capítulo 3

Tabla F.4: Catálogo de las 371 especies de mamíferos para las que se recopilaron las secuencias genéticas de las subunidades del complejo citocromo c oxidasa (COX), codificadas por mtDNA (COX 1, COX 2 y COX 3). En la tercera columna se indica el *accession number* del genoma mitocondrial de la especie.

no.	Especie	Accession number	
1	<i>Bos taurus</i>	NC_006853	28 <i>Balaenoptera omurai</i>
2	<i>Acinonyx jubatus</i>	NC_005212	29 <i>Balaenoptera physalus</i>
3	<i>Ailuropoda melanoleuca</i>	NC_009492	30 <i>Berardius bairdii</i>
4	<i>Ailurus fulgens</i>	NC_011124	31 <i>Bison bison</i>
5	<i>Ailurus fulgens styani</i>	NC_009691	32 <i>Bison bonasus</i>
6	<i>Ammotragus lervia</i>	NC_009510	33 <i>Bos grunniens</i>
7	<i>Anomalurus sp</i>	NC_009056	34 <i>Bos indicus</i>
8	<i>Antilope cervicapra</i>	NC_012098	35 <i>Bos javanicus</i>
9	<i>Aotus azarai azarai</i>	NC_018115	36 <i>Bos primigenius</i>
10	<i>Aotus nancymaae</i>	NC_018116	37 <i>Bradypus tridactylus</i>
11	<i>Apodemus agrarius</i>	NC_016428	38 <i>Bubalus bubalis</i>
12	<i>Apodemus chejuensis</i>	NC_016662	39 <i>Budorcas taxicolor</i>
13	<i>Apodemus chevrieri</i>	NC_017599	40 <i>Caenolestes fuliginosus</i>
14	<i>Apodemus peninsulae</i>	NC_016060	41 <i>Callorhinus ursinus</i>
15	<i>Arctocephalus pusillus</i>	NC_008417	42 <i>Camelus bactrianus</i>
16	<i>Arctocephalus townsendi</i>	NC_008420	43 <i>Camelus dromedarius</i>
17	<i>Arctodus simus</i>	NC_011116	44 <i>Camelus ferus</i>
18	<i>Artibeus jamaicensis</i>	NC_002009	45 <i>Canis latrans</i>
19	<i>Artibeus lituratus</i>	NC_016871	46 <i>Canis lupus</i>
20	<i>Artocephalus forsteri</i>	NC_004023	47 <i>Canis lupus chanco</i>
21	<i>Balaena mysticetus</i>	NC_005268	48 <i>Canis lupus familiaris</i>
22	<i>Balaenoptera acutorostrata</i>	NC_005271	49 <i>Canis lupus laniger</i>
23	<i>Balaenoptera bonaerensis</i>	NC_006926	50 <i>Canis lupus lupus</i>
24	<i>Balaenoptera borealis</i>	NC_006929	51 <i>Caperea marginata</i>
25	<i>Balaenoptera brydeei</i>	NC_006928	52 <i>Capra hircus</i>
26	<i>Balaenoptera edeni</i>	NC_007938	53 <i>Capricornis crispus</i>
27	<i>Balaenoptera musculus</i>	NC_001601	54 <i>Capricornis swinhoei</i>
			55 <i>Castor canadensis</i>
			56 <i>Castor fiber</i>
			57 <i>Cavia porcellus</i>
			58 <i>Cebus albifrons</i>
			59 <i>Cebus apella</i>
			60 <i>Ceratotherium simum</i>
			61 <i>Cervus elaphus</i>
			62 <i>Cervus elaphus songaricus</i>
			63 <i>Cervus elaphus xanthopygus</i>

F. Datos

no.	Especie	Accession number	no.	Especie	Accession number
64	<i>Cervus elaphus yarkandensis</i>	NC_013840	97	<i>Echinosorex gymnura</i>	NC_002808
65	<i>Cervus hortulorum</i>	NC_013834	98	<i>Echymipera rufescens australis</i>	NC_007632
66	<i>Cervus nippon centralis</i>	NC_006993	99	<i>Elaphodus cephalophorus</i>	NC_008749
67	<i>Cervus nippon kopschi</i>	NC_016178	100	<i>Elephantulus sp</i>	NC_004921
68	<i>Cervus nippon yakushimae</i>	NC_007179	101	<i>Elephas maximus</i>	NC_005129
69	<i>Cervus taiouanus</i>	NC_008462	102	<i>Enhydra lutris</i>	NC_009692
70	<i>Cervus yesoensis</i>	NC_006973	103	<i>Eospalax baileyi</i>	NC_018098
71	<i>Chalinolobus tuberculatus</i>	NC_002626	104	<i>Eothenomys chinensis</i>	NC_013571
72	<i>Chlorocebus aethiops</i>	NC_007009	105	<i>Episoriculus fumidus</i>	NC_003040
73	<i>Chlorocebus pygerythrus</i>	NC_009747	106	<i>Equus asinus</i>	NC_001788
74	<i>Chlorocebus sabaeus</i>	NC_008066	107	<i>Equus caballus</i>	NC_001640
75	<i>Chlorocebus tantalus</i>	NC_009748	108	<i>Equus hemionus</i>	NC_016061
76	<i>Choloepus didactylus</i>	NC_006924	109	<i>Eremitalpa granti</i>	NC_010304
77	<i>Chrysocloris asiatica</i>	NC_004920	110	<i>Erignathus barbatus</i>	NC_008426
78	<i>Coelodonta antiquitatis</i>	NC_012681	111	<i>Erinaceus europaeus</i>	NC_002080
79	<i>Colobus guereza</i>	NC_006901	112	<i>Eschrichtius robustus</i>	NC_005270
80	<i>Cricetulus griseus</i>	NC_007936	113	<i>Eubalaena australis</i>	NC_006930
81	<i>Crocidura russula</i>	NC_006893	114	<i>Eubalaena japonica</i>	NC_006931
82	<i>Cuon alpinus</i>	NC_013445	115	<i>Eulemur fulvus fulvus</i>	NC_012766
83	<i>Cystophora cristata</i>	NC_008427	116	<i>Eulemur fulvus mayottensis</i>	NC_012769
84	<i>Dactylopsila trivirgata</i>	NC_008134	117	<i>Eulemur macaco macaco</i>	NC_012771
85	<i>Dasyurus novemcinctus</i>	NC_001821	118	<i>Eulemur mongoz</i>	NC_010300
86	<i>Dasyurus hallucatus</i>	NC_007630	119	<i>Eumetopias jubatus</i>	NC_004030
87	<i>Daubentonia madagascariensis</i>	NC_010299	120	<i>Felis catus</i>	NC_001700
88	<i>Delphinus capensis</i>	NC_012061	121	<i>Galago senegalensis</i>	NC_012761
89	<i>Dendrohyrax dorsalis</i>	NC_010301	122	<i>Galemys pyrenaicus</i>	NC_008156
90	<i>Dicerorhinus sumatrensis</i>	NC_012684	123	<i>Galeopterus variegatus</i>	NC_004031
91	<i>Diceros bicornis</i>	NC_012682	124	<i>Giraffa camelopardalis angolensis</i>	NC_012100
92	<i>Didelphis virginiana</i>	NC_001610	125	<i>Glis glis</i>	NC_001892
93	<i>Distoechurus pennatus</i>	NC_008145	126	<i>Gorilla gorilla</i>	NC_001645
94	<i>Dromiciops gliroides</i>	NC_005826	127	<i>Gorilla gorilla gorilla</i>	NC_011120
95	<i>Dugong dugon</i>	NC_003314	128	<i>Grampus griseus</i>	NC_012062
96	<i>Echinops telfairi</i>	NC_002631	129	<i>Gulo gulo</i>	NC_009685
			130	<i>Halichoerus grypus</i>	NC_001602

F.3. Datos del capítulo 3

no.	Especie	Accession number	no.	Especie	Accession number
131	<i>Helarctos malayanus</i>	NC_009968	164	<i>Lobodon carcinophaga</i>	NC_008423
132	<i>Hemiechinus auritus</i>	NC_005033	165	<i>Loris tardigradus</i>	NC_012763
133	<i>Herpestes javanicus</i>	NC_006835	166	<i>Loxodonta africana</i>	NC_000934
134	<i>Heterocephalus glaber</i>	NC_015112	167	<i>Lutra lutra</i>	NC_011358
135	<i>Hippopotamus amphibius</i>	NC_000889	168	<i>Lynx rufus</i>	NC_014456
136	<i>Homo sapiens</i>	NC_012920	169	<i>Macaca fascicularis</i>	NC_012670
137	<i>Homo sapiens neanderthalensis</i>	NC_011137	170	<i>Macaca mulatta</i>	NC_005943
138	<i>Homo sp. Atai</i>	NC_013993	171	<i>Macaca sylvanus</i>	NC_002764
139	<i>Hydropotes inermis</i>	NC_011821	172	<i>Macaca thibetana</i>	NC_011519
140	<i>Hydropotes inermis argyropus</i>	NC_018032	173	<i>Macropus robustus</i>	NC_001794
141	<i>Hydrurga leptonyx</i>	NC_008425	174	<i>Macroscelides proboscideus</i>	NC_004026
142	<i>Hylobates agilis</i>	NC_014042	175	<i>Macrotis lagotis</i>	NC_006520
143	<i>Hylobates lar</i>	NC_002082	176	<i>Mammut americanum</i>	NC_009574
144	<i>Hylobates pileatus</i>	NC_014045	177	<i>Mammuthus columbi</i>	NC_015529
145	<i>Hylomys suillus</i>	NC_010298	178	<i>Mammuthus primigenius</i>	NC_007596
146	<i>Hyperoodon ampullatus</i>	NC_005273	179	<i>Manis pentadactyla</i>	NC_016008
147	<i>Inia geoffrensis</i>	NC_005276	180	<i>Manis tetradactyla</i>	NC_004027
148	<i>Isoodon macrourus</i>	NC_002746	181	<i>Martes flavigula</i>	NC_012141
149	<i>Jaculus jaculus</i>	NC_005314	182	<i>Martes melampus</i>	NC_009678
150	<i>Kogia breviceps</i>	NC_005272	183	<i>Martes zibellina</i>	NC_011579
151	<i>Lagenorhynchus albirostris</i>	NC_005278	184	<i>Megaptera novaeangliae</i>	NC_006927
152	<i>Lagorchestes hirsutus</i>	NC_008136	185	<i>Meles anakuma</i>	NC_009677
153	<i>Lagostrophus fasciatus</i>	NC_008447	186	<i>Meles meles</i>	NC_011125
154	<i>Lama glama</i>	NC_012102	187	<i>Melursus ursinus</i>	NC_009970
155	<i>Lama guanicoe</i>	NC_011822	188	<i>Mesocricetus auratus</i>	NC_013276
156	<i>Lasiurus borealis</i>	NC_016873	189	<i>Metachirus nudicaudatus</i>	NC_006516
157	<i>Leggadina lakedownensis</i>	NC_014696	190	<i>Microtus fortis californicus</i>	NC_015243
158	<i>Lemur catta</i>	NC_004025	191	<i>Microtus fortis fortis</i>	NC_015241
159	<i>Lepilemur hubbardorum</i>	NC_014453	192	<i>Microtus kikuchii</i>	NC_003041
160	<i>Leptonychotes weddellii</i>	NC_008424	193	<i>Microtus levis</i>	NC_008064
161	<i>Lepus capensis</i>	NC_015841	194	<i>Mirounga leonina</i>	NC_008422
162	<i>Lepus europaeus</i>	NC_004028	195	<i>Mogera wogura</i>	NC_005035
163	<i>Lipotes vexillifer</i>	NC_007629	196	<i>Monachus schauinslandi</i>	NC_008421

F. Datos

no.	Especie	Accession number	no.	Especie	Accession number
197	<i>Monodelphis domestica</i>	NC_006299	229	<i>Odobenus rosmarus ros-</i> <i>marus</i>	NC_004029
198	<i>Monodon monoceros</i>	NC_005279	230	<i>Odocoileus virginianus</i>	NC_015247
199	<i>Moschus berezovskii</i>	NC_012694	231	<i>Orcinus orca</i>	NC_014682
200	<i>Moschus moschiferus</i>	NC_013753	232	<i>Ornithorhynchus anati-</i> <i>nus</i>	NC_000891
201	<i>Muntiacus crinifrons</i>	NC_004577	233	<i>Orycterus afer</i>	NC_002078
202	<i>Muntiacus muntjak</i>	NC_004563	234	<i>Oryctolagus cuniculus</i>	NC_001913
203	<i>Muntiacus reevesi</i>	NC_004069	235	<i>Oryx dammah</i>	NC_016421
204	<i>Muntiacus reevesi mi-</i> <i>crurus</i>	NC_008491	236	<i>Oryx gazella</i>	NC_016422
205	<i>Muntiacus vuguangen-</i> <i>sis</i>	NC_016920	237	<i>Otolemur crassicauda-</i> <i>tus</i>	NC_012762
206	<i>Mus musculus</i>	NC_005089	238	<i>Ovis aries</i>	NC_001941
207	<i>Mus musculus casta-</i> <i>neus</i>	NC_012387	239	<i>Ovis canadensis</i>	NC_015889
208	<i>Mus musculus domesti-</i> <i>cus</i>	NC_006914	240	<i>Pan paniscus</i>	NC_001644
209	<i>Mus musculus molossi-</i> <i>nus</i>	NC_006915	241	<i>Pan troglodytes</i>	NC_001643
210	<i>Mus musculus musculus</i>	NC_010339	242	<i>Panthera leo persica</i>	NC_018053
211	<i>Mus terricolor</i>	NC_010650	243	<i>Panthera pardus</i>	NC_010641
212	<i>Myodes regulus</i>	NC_016427	244	<i>Panthera tigris</i>	NC_010642
213	<i>Myotis formosus</i>	NC_015828	245	<i>Panthera tigris amoyen-</i> <i>sis</i>	NC_014770
214	<i>Myrmecobius fasciatus</i>	NC_011949	246	<i>Pantholops hodgsonii</i>	NC_007441
215	<i>Mystacina tuberculata</i>	NC_006925	247	<i>Papio hamadryas</i>	NC_001992
216	<i>Naemorhedus caudatus</i>	NC_013751	248	<i>Pecari tajacu</i>	NC_012103
217	<i>Nannospalax ehrenbergi</i>	NC_005315	249	<i>Perameles gunnii</i>	NC_006521
218	<i>Nasalis larvatus</i>	NC_008216	250	<i>Perodicticus potto</i>	NC_012764
219	<i>Neodon irene</i>	NC_016055	251	<i>Petaurus breviceps</i>	NC_008135
220	<i>Neofelis nebulosa</i>	NC_008450	252	<i>Phacochoerus africanus</i>	NC_008830
221	<i>Neophoca cinerea</i>	NC_008419	253	<i>Phalanger vestitus</i>	NC_008137
222	<i>Nomascus siki</i>	NC_014051	254	<i>Phascogale tapoatafa</i>	NC_006523
223	<i>Notoryctes typhlops</i>	NC_006522	255	<i>Phascolarctos cinereus</i>	NC_008133
224	<i>Nyctereutes procyonoi-</i> <i>des</i>	NC_013700	256	<i>Phoca fasciata</i>	NC_008428
225	<i>Nycticebus coucang</i>	NC_002765	257	<i>Phoca groenlandica</i>	NC_008429
226	<i>Ochotona collaris</i>	NC_003033	258	<i>Phoca largha</i>	NC_008430
227	<i>Ochotona curzoniae</i>	NC_011029	259	<i>Phoca vitulina</i>	NC_001325
228	<i>Ochotona princeps</i>	NC_005358	260	<i>Phocarctos hookeri</i>	NC_008418
			261	<i>Phocoena phocoena</i>	NC_005280
			262	<i>Physeter catodon</i>	NC_002503
			263	<i>Piliocolobus badius</i>	NC_008219

F.3. Datos del capítulo 3

no.	Especie	Accession number	no.	Especie	Accession number
264	<i>Pipistrellus abramus</i>	NC_005436	296	<i>Rattus leucopus</i>	NC_014855
265	<i>Platanista minor</i>	NC_005275	297	<i>Rattus lutreolus</i>	NC_014858
266	<i>Plecotus auritus</i>	NC_015484	298	<i>Rattus norvegicus</i>	NC_001665
267	<i>Plecotus rafinesquii</i>	NC_016872	299	<i>Rattus praetor</i>	NC_012461
268	<i>Pongo abelii</i>	NC_002083	300	<i>Rattus rattus</i>	NC_012374
269	<i>Pongo pygmaeus</i>	NC_001646	301	<i>Rattus sordidus</i>	NC_014871
270	<i>Pontoporia blainvilliei</i>	NC_005277	302	<i>Rattus tanzeumi</i>	NC_011638
271	<i>Potorous tridactylus</i>	NC_006524	303	<i>Rattus tunneyi</i>	NC_014861
272	<i>Presbytis melalophos</i>	NC_008217	304	<i>Rattus villossissimus</i>	NC_014864
273	<i>Prionailurus bengalensis euptilurus</i>	NC_016189	305	<i>Rhinoceros sondaicus</i>	NC_012683
274	<i>Procavia przewalskii</i>	NC_014875	306	<i>Rhinoceros unicornis</i>	NC_001779
275	<i>Procyon lotor</i>	NC_004919	307	<i>Rhinolophus ferrumequinum korai</i>	NC_016191
276	<i>Proedromys liangshannensis</i>	NC_009126	308	<i>Rhinolophus formosae</i>	NC_011304
277	<i>Propithecus coquereli</i>	NC_011053	309	<i>Rhinolophus monoceros</i>	NC_005433
278	<i>Przewalskium albirostris</i>	NC_016707	310	<i>Rhinolophus pumilus</i>	NC_005434
279	<i>Pseudocheirus peregrinus</i>	NC_006519	311	<i>Rhinopithecus avunculus</i>	NC_015485
280	<i>Pseudois schaeferi</i>	NC_016689	312	<i>Rhinopithecus bieti</i>	NC_015486
281	<i>Pseudomys chapmani</i>	NC_014698	313	<i>Rhinopithecus bieti</i> 1	NC_018058
282	<i>Pteropus dasymallus</i>	NC_002612	314	<i>Rhinopithecus bieti</i> 2	NC_018060
283	<i>Pteropus scapulatus</i>	NC_002619	315	<i>Rhinopithecus brelichi</i>	NC_018057
284	<i>Puma concolor</i>	NC_016470	316	<i>Rhinopithecus roxellana</i>	NC_008218
285	<i>Pusa caspica</i>	NC_008431	317	<i>Rhinopithecus strykeri</i>	NC_018059
286	<i>Pusa hispida</i>	NC_008433	318	<i>Rhyncholestes raphanurus</i>	NC_005829
287	<i>Pusa sibirica</i>	NC_008432	319	<i>Rousettus aegyptiacus</i>	NC_007393
288	<i>Pygathrix cinerea</i> 2 RL-2012	NC_018063	320	<i>Rucervus eldi</i>	NC_014701
289	<i>Pygathrix cinerea</i> 1 RL-2012	NC_018062	321	<i>Rusa unicolor swinhoei</i>	NC_008414
290	<i>Saimiri boliviensis boliviensis</i>	NC_008220	322	<i>Saimiri sciureus</i>	NC_018096
291	<i>Sciurus vulgaris</i>	NC_018061	323	<i>Sciurus vulgaris</i>	NC_012775
292	<i>Rangifer tarandus</i>	NC_007703	324	<i>Semnopithecus entellus</i>	NC_002369
293	<i>Rattus exulans</i>	NC_012389	325	<i>Sminthopsis crassicaudata</i>	NC_008215
294	<i>Rattus fuscipes</i>	NC_014867	326		NC_007631

F. Datos

no.	Especie	Accession number	no.	Especie	Accession number
327	<i>Sminthopsis douglasi</i>	NC_006517	361	<i>Ursus thibetanus formosanus</i>	NC_009331
328	<i>Sorex unguiculatus</i>	NC_005435	362	<i>Ursus thibetanus mupinensis</i>	NC_008753
329	<i>Sousa chinensis</i>	NC_012057	363	<i>Ursus thibetanus thibetanus</i>	NC_011118
330	<i>Spilogale putorius</i>	NC_010497	364	<i>Ursus thibetanus ussuricus</i>	NC_011117
331	<i>Stenella attenuata</i>	NC_012051	365	<i>Varecia variegata variegata</i>	NC_012773
332	<i>Stenella coeruleoalba</i>	NC_012053	366	<i>Vicugna pacos</i>	NC_002504
333	<i>Sus scrofa</i>	NC_000845	367	<i>Vicugna vicugna</i>	NC_013558
334	<i>Sus scrofa domesticus</i>	NC_012095	368	<i>Vombatus ursinus</i>	NC_003322
335	<i>Sus scrofa taiwanensis</i>	NC_014692	369	<i>Vulpes vulpes</i>	NC_008434
336	<i>Sympthalangus syndactylus</i>	NC_014047	370	<i>Zaglossus bruijni</i>	NC_006364
337	<i>Tachyglossus aculeatus</i>	NC_003321	371	<i>Zalophus californianus</i>	NC_008416
338	<i>Talpa europaea</i>	NC_002391			
339	<i>Tamandua tetradactyla</i>	NC_004032			
340	<i>Tarsipes rostratus</i>	NC_006518			
341	<i>Tarsius bancanus</i>	NC_002811			
342	<i>Tarsius syrichta</i>				
343	<i>Thryonomys swinderianus</i>				
344	<i>Thylacinus cynocephalus</i>				
345	<i>Thylamys elegans</i>				
346	<i>Trachypithecus obscurus</i>				
347	<i>Tremarctos ornatus</i>				
348	<i>Trichechus manatus</i>				
349	<i>Trichosurus vulpecula</i>				
350	<i>Tscherskia triton</i>				
351	<i>Tupaia belangeri</i>				
352	<i>Tursiops aduncus</i>				
353	<i>Tursiops truncatus</i>				
354	<i>Uncia uncia</i>				
355	<i>Urotrichus talpoides</i>				
356	<i>Ursus americanus</i>				
357	<i>Ursus arctos</i>				
358	<i>Ursus maritimus</i>				
359	<i>Ursus spelaeus</i>				
360	<i>Ursus thibetanus</i>				

F.3. Datos del capítulo 3

Tabla F.5: Catálogo de secuencias de nDNA que codifican para las 10 subunidades pertenecientes al complejo citocromo c oxidasa.

no.	Especie	COX 4	COX 5A	COX 5B	COX 6A2
1	<i>Ailuropoda melanoleuca</i>	XM_002913384.1	XM_002922971.1	XM_002912393.1	XM_002925805.1
2	<i>Ateles belzebuth</i>	-	-	-	-
3	<i>Bos taurus</i>	NM_001001439	NM_001002891.1	NM_001034046.2	NM_174522.2
4	<i>Callicebus donacophilus</i>	-	DQ987248.1	-	-
5	<i>Callithrix jacchus</i>	-	XM_002753347.2	-	XM_002761936.2
6	<i>Callithrix pygmaea</i>	-	DQ987246.1	-	-
7	<i>Canis lupus familiaris</i>	XM_536759.3	XM_535544.3	NM_001144130.1	XM_851113.2
8	<i>Cavia porcellus</i>	XM_003461023.1	XM_003463579.1	XM_003471692.1	XM_003477896.1
9	<i>Cricetulus griseus</i>	XM_003495107.1	-	XM_003498291.1	NM_0035116063.1
10	<i>Colobus guereza</i>	-	DQ987244.1	-	-
11	<i>Equus caballus</i>	XM_001502557.1	XM_001491856.3	XM_001493723.2	NM_001500332.1
12	<i>Eulemur fulvus</i>	-	DQ987251.1	-	-
13	<i>Gorilla gorilla</i>	-	DQ987240.1	-	-
14	<i>Homo sapiens</i>	NM_001861.3	NM_004255.3	CR541727.1	M83308.1
15	<i>Loxodonta africana</i>	XM_003418076.1	XR_133810.1	XM_003422005.1	NM_003418792.1
16	<i>Macaca fascicularis</i>	-	-	-	-
17	<i>Macaca mulatta</i>	NM_001193548.1	NM_001040279.1	XM_001098868.2	NM_001194578.1
18	<i>Macaca silenus</i>	-	-	-	-
19	<i>Monodelphis domestica</i>	-	-	-	-
20	<i>Mus musculus</i>	BC132269.1	NM_007747.2	NM_009942.2	U08439.1
21	<i>Nomascus gabriellae</i>	-	DQ987242.1	-	-
22	<i>Nomascus leucogenys</i>	NM_003272506.1	XM_003267228.1	XM_003282429.1	NM_003280466.1
23	<i>Nycticebus couang</i>	-	DQ987249.1	-	-
24	<i>Oryctolagus cuniculus</i>	NM_001170882.1	-	XM_002710006.1	NM_002721729.1
25	<i>Otolemur crassicaudatus</i>	-	DQ987250.1	-	-
26	<i>Pan paniscus</i>	-	DQ987239.1	-	XM_001158801.2
27	<i>Pan troglodytes</i>	NM_001251915.1	NM_001118913.1	XM_001154903.2	-

F. Datos

no.	Especie	COX 4	COX 5A	COX 5B	COX 6A2
28	<i>Papio anubis</i>	-	DQ987245.1	-	-
29	<i>Pongo abelii</i>	NM_001133005.1	XM_002824314.1	NM_001131385.1	NM_001131893.1
30	<i>Pongo pygmaeus</i>	-	DQ987241.1	-	-
31	<i>Rattus norvegicus</i>	NM_017202.1	NM_145783.1	BC083179.1	NM_012812.3
32	<i>Rousettus leschenaultii</i>	GU292805.1	-	-	-
33	<i>Saguinus labiatus</i>	-	DQ987247.1	-	-
34	<i>Saimiri sciureus</i>	-	AY585857.1	-	-
35	<i>Sus scrofa</i>	AK399746.1	XM_003482240.1	AK400056.1	XM_003481013.1
36	<i>Symphalangus syndactylus</i>	-	DQ987243.1	-	-
37	<i>Tarsius syrichta</i>	-	AY236506.1	-	-
38	<i>Trachypithecus cristatus</i>	-	-	-	-

Tabla F.6: Continuación de la tabla F.5.

no.	Especie	COX 6B1	COX 6C	COX 7A1
1	<i>Ailuropoda melanoleuca</i>	HQ380463.1	XR_097052.1	XM_002920932.1
2	<i>Ateles belzebuth</i>	-	-	-
3	<i>Bos taurus</i>	NM_176675.3	XM_002690379.1	BC114907.1
4	<i>Callicebus donacophilus</i>	-	-	-
5	<i>Callithrix jacchus</i>	XM_002762021.2	XM_002759357.2	XM_002762060.2
6	<i>Callithrix pygmaea</i>	-	-	-
7	<i>Canis lupus familiaris</i>	XM_003432596.1	XM_850997.2	-
8	<i>Cavia porcellus</i>	XM_003467210.1	XM_003479873.1	XM_003467228.1
9	<i>Cricetus griseus</i>	XM_003505555.1	-	-
10	<i>Colobus guereza</i>	-	-	-
11	<i>Equus caballus</i>	XM_001492385.3	XM_001492004.3	XM_001492779.2
12	<i>Eulemur fulvus</i>	-	-	-
13	<i>Gorilla gorilla</i>	-	-	-
14	<i>Homo sapiens</i>	NG_012193.1	BC000187.2	AK312091.1

F.3. Datos del capítulo 3

no.	Especie	COX 6B1	COX 6C	COX 7A1
15	<i>Loxodonta africana</i>	XM_003420756.1	XM_003408413.1	XM_003422300.1
16	<i>Macaca fascicularis</i>	AB179393.1	-	-
17	<i>Macaca mulatta</i>	NM_001040281.1	XM_001097441.2	NM_001040278.1
18	<i>Macaca sylvanus</i>	-	AY236508.1	-
19	<i>Monodelphis domestica</i>	XM_001363836.1	-	-
20	<i>Mus musculus</i>	NM_025628.2	NM_053071.2	BC060974.1
21	<i>Nomascus gabriellae</i>	-	-	-
22	<i>Nomascus leucogenys</i>	XM_003280044.1	XM_003262020.1	-
23	<i>Nycticebus coucang</i>	-	AY236512.1	-
24	<i>Oryctolagus cuniculus</i>	XM_002722221.1	XM_002721825.1	XM_002722250.1
25	<i>Otolemur crassicaudatus</i>	-	-	-
26	<i>Pan paniscus</i>	-	-	-
27	<i>Pan troglodytes</i>	XM_001160605.1	XM_001151309.2	XM_003316334.1
28	<i>Papio anubis</i>	-	-	-
29	<i>Pongo abelii</i>	NM_001131741.1	XM_002834759.1	-
30	<i>Pongo pygmaeus</i>	-	-	-
31	<i>Rattus norvegicus</i>	NG_028330.1	BC058480.1	XM_001075627.2
32	<i>Rousettus leschenaultii</i>	-	-	-
33	<i>Saguinus labiatus</i>	-	-	-
34	<i>Saimiri sciureus</i>	-	-	-
35	<i>Sus scrofa</i>	NM_001097497.1	AK392221.1	NM_214411.1
36	<i>Symphalangus syndactylus</i>	-	-	-
37	<i>Tarsius syrichta</i>	AY236504.1	AY236503.1	AY585864.1
38	<i>Trachypithecus cristatus</i>	-	AY236509.1	-

Tabla F.7: Continuación de la tabla F.5.

no.	Especie	COX 7B	COX 7C	COX 8C
1	<i>Ailuropoda melanoleuca</i>	XM_002930510.1	XM_002913621.1	XM_002930619.1

F. Datos

no.	Especie	COX 7B	COX 7C	COX 8C
2	<i>Atelos belzebuth</i>	-	-	AA910507.1
3	<i>Bos taurus</i>	NM_175795.3	NM_175831.3	NM_001114517.2
4	<i>Callicebus donacophilus</i>	-	-	-
5	<i>Callithrix jacchus</i>	XM_003735770.1	XM_002744779.2	XM_002755707.1
6	<i>Callithrix pygmaea</i>	-	-	-
7	<i>Canis lupus familiaris</i>	XM_003435588.1	XM_847363.2	XM_003639656.1
8	<i>Canis porcellus</i>	XM_003470974.1	XM_003479466.1	XM_003461297.1
9	<i>Cricetulus griseus</i>	-	-	XM_003515383.1
10	<i>Colobus guereza</i>	-	-	-
11	<i>Equus caballus</i>	XM_003365825.1	XM_003362849.1	XM_003362667.1
12	<i>Eulemur fulvus</i>	-	-	AY254828.1
13	<i>Gorilla gorilla</i>	-	-	-
14	<i>Homo sapiens</i>	AK311879.1	BC007498.2	-
15	<i>Loxodonta africana</i>	XM_003412712.1	-	-
16	<i>Macaca fascicularis</i>	-	-	-
17	<i>Macaca mulatta</i>	NM_001258140.1	XM_001081984.2	-
18	<i>Macaca silenus</i>	-	-	-
19	<i>Monodelphis domestica</i>	-	-	-
20	<i>Mus musculus</i>	BC024350.1	XM_003689339.1	BC086930.1
21	<i>Nomascus gabriellae</i>	-	-	-
22	<i>Nomascus leucogenys</i>	XM_003282497.1	XM_003261573.1	-
23	<i>Nycticebus coucang</i>	-	-	-
24	<i>Oryctolagus cuniculus</i>	XM_002722477.1	XM_002721439.1	XM_002724105.1
25	<i>Otolemur crassicaudatus</i>	-	-	-
26	<i>Pan paniscus</i>	-	-	-
27	<i>Pan troglodytes</i>	XR_129485.1	AK307009.1	-
28	<i>Papio anubis</i>	-	-	-
29	<i>Pongo abelii</i>	XR_092709.1	XM_002809590.1	-
30	<i>Pongo pygmaeus</i>	-	-	-

F.3. Datos del capítulo 3

no.	Especie	COX 7B	COX 7C	COX 8C
31	<i>Rattus norvegicus</i>	FQ213156.1	FQ224715.1	FQ223790.1
32	<i>Rousettus leschenaultii</i>	-	-	-
33	<i>Saguinus labiatus</i>	-	-	-
34	<i>Saimiri sciureus</i>	-	AY585860.1	-
35	<i>Sus scrofa</i>	AK392166.1	DQ629155.1	NM_001097500.1
36	<i>Symphalangus syndactylus</i>	-	-	-
37	<i>Tarsius syrichta</i>	-	AY236505.1	AY254827.1
38	<i>Trachypithecus cristatus</i>	-	-	-

F. Datos

F.4 Información estructural

Para los trabajos expuestos en los capítulos 2 y 3, se realizaron análisis sobre los modelos atómicos como 1be3 (complejo III) y 2occ (complejo IV) en *Protein Data Bank* (PDB). En la tabla F.8 se resume alguna información de interés sobre sendos modelos, obtenida de la misma base de datos:

Tabla F.8: Modelos atómicos analizados en esta tesis.

Id.	Especie	Técnica	Resolución	Referencia
1be3	<i>Bos taurus</i>	Dif. rayos X Enlace: http://bit.ly/1566Fb0	3.00Å	(?)
2occ	<i>Bos taurus</i>	Dif. rayos X Enlace: http://bit.ly/1aBKbGU	2.30Å	(?)

F.5 Edad de las proteínas del complejo IV

En algunos análisis llevados a cabo en el capítulo 3, se computaron las edades (en millones de años) de las proteínas codificadas por nDNA, presentes en el complejo citocromo c oxidasa. Estos datos fueron obtenidos mediante el servidor *ProteinHistorian* (?), cuya interfaz web está disponible en: <http://lighthouse.ucsf.edu/ProteinHistorian/>. En la tabla F.9 se muestra dicha información.

F.5. Edad de las proteínas del complejo IV

Tabla F.9: Edad de las cadenas del complejo IV codificadas por nDNA de *Bos taurus*.

Subunidad	Edad (mill. años)	UniProtKB [§]	Taxón de origen
COX 1	4200	P00396	Bacteria
COX 2	4200	P68530	Bacteria
(*) COX 3	4200	P00415	Bacteria
COX 4A	1368	P00423	Opisthokonta
COX 5A	1368	P00426	Opisthokonta
COX 5B	1628	P00428	Eukaryota
(†) COX 6A2	??	P00471	??
COX 6B1	1368	P00429	Opisthokonta
COX 6C	842.0	P04038	Deuterostomia
COX 7A1	176.1	P07470	Theria
COX 7B	361.2	P13183	Tetrapoda
COX 7C	454.6	P00430	Euteleostomi
(*) COX 8B	910	P10175	Bilateria

§ Identificador de proteína en UniProt (<http://www.uniprot.org/>), el cual es el dato de entrada del servidor *ProteinHistorian*.

†La edad de la subunidad COX 6A2 fue imposible determinar por *ProteinHistorian* por razones desconocidas.

* La edad de COX 3 y COX 8B fue estimada mediante un algoritmo distinto (*Dollo parsimony*) al que se usó para estimar la edad del resto de las subunidades (*Wagner parsimony*). Véase <http://lighthouse.ucsf.edu/ProteinHistorian/methods.html> para más detalles sobre los métodos de *ProteinHistorian*.

F. Datos
