Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster Universitario en
Ingeniería Computacional y Sistemas Inteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila -
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A I C S I

Tesis de Máster

Comparación de haplotipos mitocondriales en presencia de incertidumbre: Aplicación a la búsqueda de linajes maternos vascos

David Ignacio Salinas Fernández de Landa

Tutor(a/es)

Borja Calvo Molinos
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática

Aritz Perez Martinez
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática

informatika fakultatea
facultad de informática

KZAA /CCIA

Septiembre 2014

Índice

1. Introducción

A lo largo de la historia se han dado varios eventos migratorios de la población vasca, fundamentalmente con destino América (USA, Argentina, Canada, ...). En ciertas regiones de estos países existen amplias colonias de descendientes de vascos, muy interesados en saber más sobre sus orígenes.

A la hora de estudiar la dinámica de las poblaciones habitualmente se recurre al uso del ADN mitrocondrial. Este tiene dos características que lo hacen interesante, su estabilidad en la población y que es heredado de forma exclusiva por vía materna (ver Figura 1). Por este motivo, el perfil de ADN mitocondrial o haplotipo mitocondrial es una manera sencilla de investigar la ascendencia materna de las personas.

El objetivo de este proyecto es desarrollar una aplicación Web que permita realizar búsquedas de perfiles en una base de datos de ADN mitocondrial. La principal dificultad es la búsqueda en presencia de incertidumbre, problema que será abordado en este proyecto.

La base de datos de perfiles ha sido obtenida a lo largo de los años por el Grupo de Investigación Consolidado BIOMICS, a través de sus diferentes proyectos de investigación y gracias a la donación que las personas han hecho de sus muestras. La razón de ser de este proyecto es el de devolver a estas personas y a la población en general un poco de lo obtenido. Por este motivo se ha desarrollado una aplicación Web que permite a cualquier persona que disponga de su perfil genético de ADN mitocondrial buscar en la base de datos (que actualmente cuenta con más de 1.000 perfiles) para saber en qué regiones del País Vasco y/o América hay personas pertenecientes a su mismo linaje materno.

1.1. Introducción a la genética

Todos hemos escuchado alguna vez que es nuestro ADN lo que nos hace únicos. Esta molécula no solo nos diferencia de otros seres vivos, también está relacionada con como somos. Lo que no es tán conocido es que en nuestro organismo hay dos tipos de ADN, el ADN nuclear y el ADN mitocondrial. En este proyecto nos centraremos en este último, pero antes de ver que es y en que consiste veremos brevemente qué es el ADN.

Ya a mediados del siglo XIX un monje austríaco, Gregor J. Mendel, definió la manera en la que las

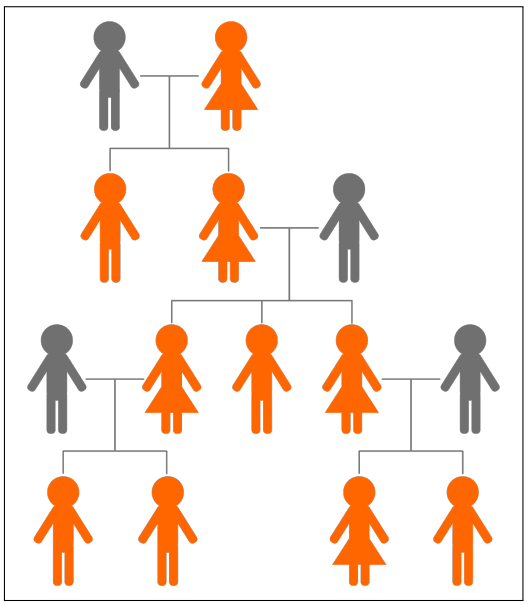

Figura 1: Herencia de ADN mitocondrial

características físicas (también denominadas fenotipo) son transmitidas de generación en generación. Las leyes de Medel supusieron una revolución científica, pero no fue más que el comienzo. Tan solo cuatro años después de que Mendel propusiese sus conocidas leyes, un médico y biólogo suizo, Friederich Miescher, aisló por primera vez el ADN, una sustancia que se encontraba en el núcleo de las células y que llamó nucleina. Sin embargo, no fue hasta casi un siglo después que ambos hechos (la herencia y el ADN) se llegaron a relacionar.

1.1.1. Estructura del ADN

En 1953, los biólogos moleculares James D. Watson y Francis H. C. Crick, estadounidense el primero e inglés el segundo, propusieron el modelo actual de la estructura del ADN[1]. Según este modelo, el ADN está formado por la famosa doble hélice. En ella, cada una de las dos hebras está formada por una secuencia de nucleótidos unidos entre sí por enlaces fosfodiester. Esta secuencia, que en términos informáticos se puede ver como un string, está formada por tan solo cuatro tipos de nucleótidos: la adenina (A), la timinia (T), la citosina (C) y la guanina (G). En todos los casos, los nucleótidos están formados por un monosacárido,

[1]Es importante remarcar que no todas las moléculas de ADN tienen la estructura descrita por Watson y Crick. Un ejemplo es el DN de los retroviruses, como el del SIDA, que sólo está formado por una única cadena

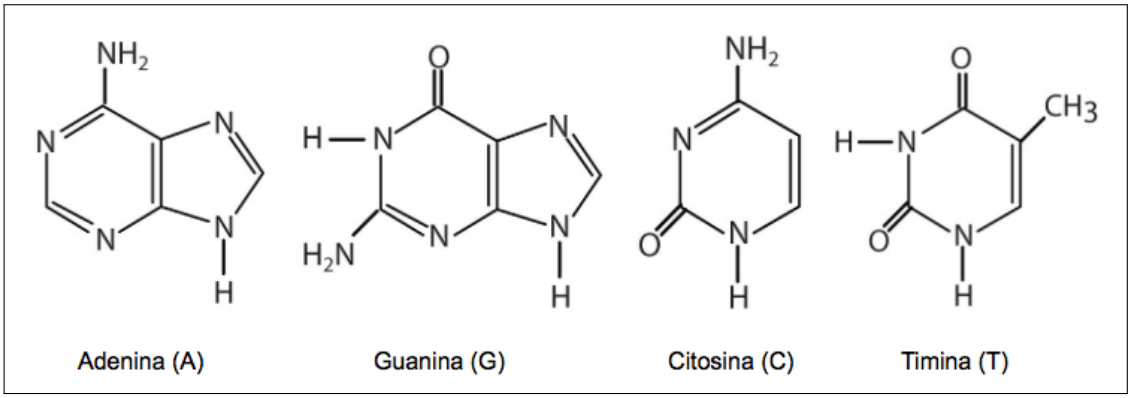

Figura 2: Bases nitrogenadas que forman parte de los nucleótidos del ADN.

la desoxiribosa, un grupo fosfato y una base nitrogenada. Es esta última la que marca la diferencia entre los cuatro nucleótidos (ver Figura 2).

Como ya se ha mencionado, el ADN se encuentra formando una doble hélice. Esto es posible debido a que los nucleótidos son complementarios dos a dos. Cuando una adenina y una timina se enfrentan, entre ellas se establecen dos enlaces de tipo puente de hidrógeno. Por su parte, cuando la citosina se enfrenta a la guanina, se establecen tres puentes de hidrógeno que estabilizan la unión. De esta manera, se dice que la adenina es la base complementaria de la timina (y viceversa) y que la guanina es la complementaria de la citosina. Esta complementariedad es fundamental en la replicación del ADN, es decir, en la generación de nuevas moléculas de ADN tomando como molde una molécula existente (copiando la secuencia correspondiente a la molécula molde).

El genoma humano esta formado por algo más de 3.000 millones de pares bases repartidas en 23 cromosomas. No solo eso, el ser humano es diploide, lo que significa que en todas las células (excepto los gametos) existen dos copias completas del genoma. Esto supone que en el núcleo de cada una de las células[2] de nuestro organismo existe ADN por ?valor? de más de 6.000 millones de pares de nucleótidos. Toda esta cantidad de material se encuentra organizada y estructurada en lo que se denomina cromatina, que es la manera en la que, por medio de una serie de proteínas, las histonas, las moléculas de ADN adquieren una estructura compacta.

Es de sobra conocido que, salvo que se trate de gemelos, dos individuos no tienen la misma secuencia

[2]Los glóbulos rojos son una excepción, ya que carecen de núcleo.

de ADN. Existe distintos tipos de variabilidad genética, aunque la más estudiada son los llamados SNPs (single nucleotide polymorphisms, polimorfismos de un nucleótido). Como su nombre sugiere, los SNPs implican diferencias en una única posición del genoma. A este respecto cabe destacar que para que una posición se considere un SNP es necesario que hay dos o más variantes que estén ampliamente representadas en la población.

1.1.2. El ADN como almacén de información

En términos de codificación de información, podemos ver el ADN como una secuencia, un string formado por un alfabeto de cuatro letras. Ahora bien, ¿en qué consiste la información que se almacena en ese string?. Para entender esto tenemos que centrar nuestra atención en otro tipo de molécula presente en nuestras células: las proteínas.

Las proteínas, al igual que ADN, son moléculas poliméricas formadas por una secuencia de unidades, los aminoácidos. Pese a ser ambos tipos de moléculas cadenas de unidades monoméricas, hay una diferencia fundamental; mientras que solo hay cuatro nucleótidos básicos en el ADN, existen algo más de 21 aminoácidos básicos en los seres humanos[3]. Por este motivo, la variedad de moléculas que se puede formar combinando estas unidades fundamentales es considerablemente mayor.

Es precisamente esta gran heterogeneidad, asociada fundamentalmente a la estructura tridimensional que las cadenas de aminoácidos adquieren en su medio, la que posibilita que las proteínas desempeñen un gran número de tareas, desde actividad enzimática, regulando las reacciones químicas que se dan en nuestro organismo, hasta funciones estructurales, de comunicación, etc. De hecho, hoy en día se estima que en el ser humano hay entre 250.000 y 1.000.000 de proteínas diferentes.

Las proteínas, con su gran heterogeneidad, son las encargadas de desempeñar la gran mayoría de las tareas necesarias para la vida pero, ¿cómo es capaz el organismo de fabricar todas las proteínas necesarias? Aquí es donde entra el papel fundamental del ADN, ya que es esta molécula la que contiene, entre otras cosas, la información necesaria para sintetizar las proteínas. Esta información está organizada en genes, fragmentos de secuencia que codifican proteínas.

[3]En total se conocen más de 500 aminoácidos, aunque solo unos pocos de ellos forman habitualmente parte de las proteínas.

1.1.3. El ADN mitocondrial y genética de poblaciones

Las células de nuestro cuerpo contienen en su interior distintos tipos de estructuras, cada una de las cuales desempeña una función. Entre estas está el núcleo, lugar donde se localizan los cromosomas. Fuera del núcleo se encuentran las llamadas mitocondrias, las cuales juegan un papel esencial en el metabolismo energético de la célula.

No está claro cual es el origen de este orgánulo, aunque algunas hipótesis apuntan a que, en tiempos remotos, se trataría de células independientes que, por un proceso de simbiosis, entraron a formar parte de células más complejas. El hecho es que, fuese de esta manera o de otra, estas estructuras tienen su propio ADN en el cual se codifican unas pocas proteínas implicadas en el metabolismo energético.

Comparado con el ADN nuclear, que está formado por más de 3.000 millones de pares de bases, el ADN mitocondrial (ADNmt) es extremadamente pequeño (16.568 pares de bases). Además, se trata de una molécula circular, esto es, el nucleótido 16.568 se une directamente al nucleótido 1.

El ADNmt tiene dos características que lo hacen especialmente interesante en el campo de la genética de poblaciones. Por una parte, a diferencia de lo que ocurre con el ADN nuclear, la herencia es exclusivamente por vía materna. Por otra parte, la aparición de nuevos SNPs en el ADNmt es un evento extremadamente raro, por lo que la secuencia permanece inalterada durante generaciones.

Estas dos características hacen que el análisis del ADNmt sea una buena forma de estudiar los linajes maternos, permitiendo determinar relaciones de parentesco vía materna; a escala más global, esta molécula permite a los investigadores estudiar la dinámica de las poblaciones humanas.

Hoy en día, gracias al avance de las tecnologías de ultrasecuenciación, es posible resecuenciar el ADNmt completo para cada inviduo que va a ser analizado. Sin embargo, habitualmente la determinación del perfil de ADNmt se realiza por técnicas de secuenciación clásica, cuya capacidad de secuenciación se reduce a unos pocos cientos de bases. Por este motivo, es habitual que el análisis se limite tan solo a algunas regiones que presentan una mayor variabilidad en la población.

Hay definidas, fundamentalmente, 2 regiones hipervariables, la HVRI (de la base 16024 a la 16364) y la HVRII (de la base 73 a la 340); más recientemente es habitual la resecuenciación de la región de control completa (de la base 16024 a la 576)[4]. Dada la baja variabilidad del ADNmt, en lugar de recoger

[4]Hay que recordar que el ADNmt es circular, por lo que la siguiente base a la última es la primera.

Pos.	1				5					10					15					20					25
Ref.	A	T	T	G	C	C	G	T	C	A	T	G	T	C	T	G	C	A	C	A	G	T	A	G	C
H1	A	T	T	G	C	C	G	T	A	A	T	G	T	C	A	G	C	A	C	A	G	T	A	G	C
H2	A	T	T	G	C	C	A	T	C	A	T	G	T	C	T	G	C	A	C	C	G	T	A	G	C

Figura 3: Ejemplo de valor de referencia y dos haplotipos.

en el perfil genético toda la información, únicamente se anotan los cambios con respecto a una referencia común, que habitualmente es la llamada Revised Cambridge Reference Sequence (rCRS) [3].

1.2. Haplotipos y Haplogrupos

Como se ha visto anteriormente, en este proyecto se trabaja con haplotipos, pero ¿qué es un haplotipo?. Los haplotipos son una combinación de alelos de diferentes locus de un cromosoma (o, en nuestro caso, del ADNmt) que son transmitidos juntos. Un haplotipo puede ser un locus, varios loci o un cromosoma entero dependiendo del número de recombinación que ha ocurrido entre un conjunto dado de loci.

A efectos prácticos, para este proyecto los haplotipos son cadenas de números y letras. Indicando de esta forma aquellas posiciones del ADNmt cuyo valor es diferente en la muestra respecto del valor de referencia [3]. Veamos esto con un ejemplo:

De la Figura 3 se obtiene que los haplotipos H1 y H2 quedarán de la siguiente manera:

- H1: 9A,15A
- H2: 7A, 20C

Un haplogrupo es un grupo grande de haplotipos. Estos haplogrupos incluyen a personas con perfiles genéticos similares que comparten un antepasado común.

Los haplogrupos más antiguos son más grandes y de ellos descienden numerosos subgrupos. Se puede visualizar los haplogrupos mitocondriales como un árbol filogenético, tal y como se muestra en la Figura 3.

La distribución geográfica actual de los haplogrupos, junto con su árbol filogenético, permiten estudiar

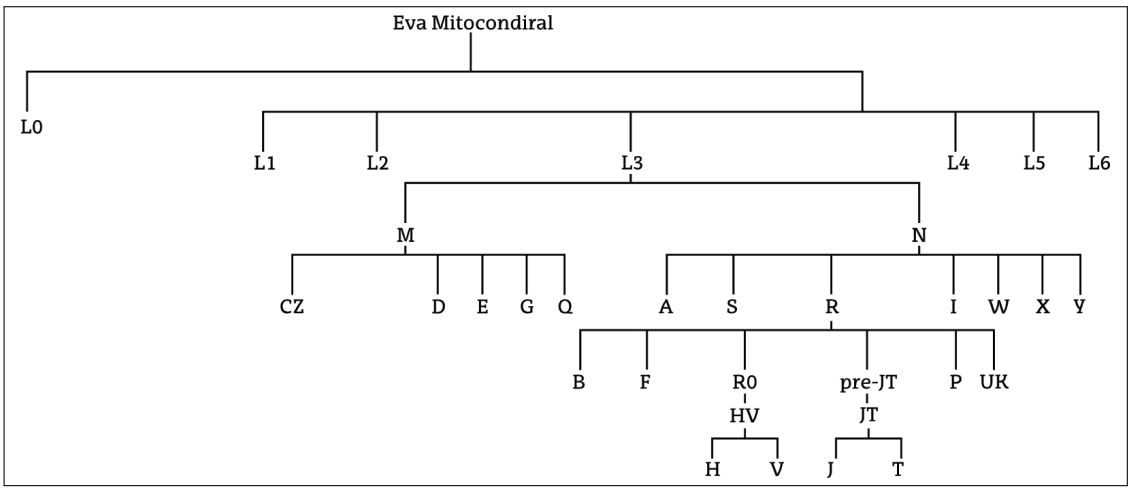

Figura 4: Versión reducida del árbol de haplogrupos.

cómo se han desplazado por el planeta los diferentes grupos demográficos (ver Figura 4).

1.3. Introducción a la Teoría de la Información

Emplearemos la Teoría de la Información [ref 1] para analizar las muestras disponibles (haplotipos de la base de datos) y poder establecer ciertas suposiciones razonables que nos permitan proponer una medida de confianza de la correspondencia entre dos haplotipos cuando contienen valores no observados en alguna posición (ver apartado Cálculo del nivel de confianza en la comparación de haplotipos). Para ello trataremos las diferentes posiciones de los haplotipos $i \in 1..,16568$ como variables aleatorias X_i que toman los estados $xi \in C, G, T, A$. Cada variable aleatoria X_i estará distribuida conforme a la distribución p_i. En este proyecto usaremos la distribución empírica que obtendremos de los N haplotipos disponibles por medio de estimadores maximoverosímiles.

1.3.1. Entropía

La entropía de una variable aleatoria X_i distribuida conforme a p_i se define como:

$$H(X_i) = -\sum_{x_i} p_i(x_i) log_2 p_i(x_i)$$

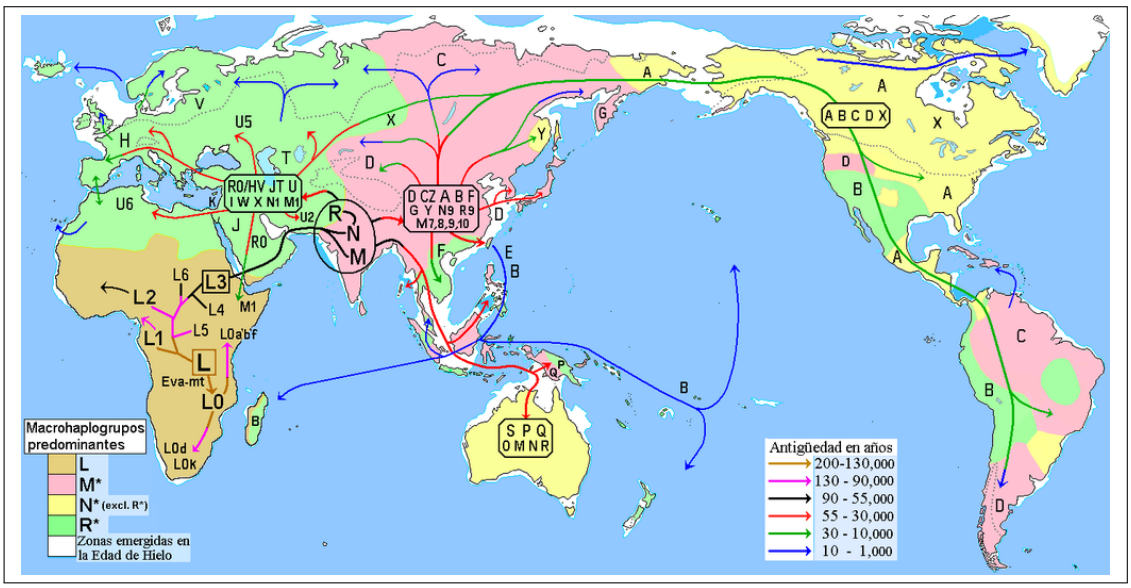

Figura 5: Migraciones humanas en haplogrupos mitocondriales.

Esta cantidad cuantifica la incertidumbre que envuelve a la variable aleatoria X_i y es una medida de la uniformidad de su distribución p_i. Por ejemplo, el mayor valor de la entropía se obtiene cuando X_i esta distribuida uniformemente, e.g. X_i con $x_i \in C, G, T, A$ con $p_i(C) = p_i(G) = p_i(T) = p_i(A) = 0{,}25$ tiene un valor de entropía máximo $H(X_i) = -log\frac{1}{4}$. Por otra parte el menor valor de la entropía, $H(X_i) = 0$, se obtiene cuando un estado tiene probabilidad uno, $max(P(X_i)) = 1$; esto es cuando tenemos certeza absoluta sobre el valor que tomará la variable aleatoria.

1.3.2. Información mutua

La información mutua entre dos variables aleatorias X_i y X_j se define en términos de la entropía como:

$$\begin{aligned} I(X_i, X_j) &= H(X_i) + H(H_j) - H(X_i, X_j) \\ &= \sum_{x_i, x_j} p_{i,j}(x_i, x_j) \log \frac{p_{i,j}(x_i, x_j)}{p_i(x_i) \cdot p_j(x_j)} \end{aligned} \tag{1}$$

siendo $p_{i,j}$ la distribución conjunta de las variables X_i y X_j. La información muytua se puede interpretar como la reducción en la incertidumbre de la variable X_i cuando se conoce el estado que toma la

variable X_j, y viceversa. Además, la información mutua entre X_i y X_j mide la fuerza de la dependencia (incondicional) entre ambas variables. Es decir, cuanto mayor es el valor de $I(X_i, X_j)$ mas dependencia muestran las variables X_i y X_j.

1.3.3. Test de la independencia

En la literatura [2] se ha propuesto un test que permite decidir si dos variables aleatorias se pueden considerar dependientes basado en la distribución chi cuadrado χ^2. El test se basa en que el estadístico $2 \cdot N \cdot I(X_i, X_j)$ sigue (asintóticamente) una distribución χ^2 con $d = (r_i - 1) \cdot (r_j - 1)$ grados de libertad cuando X_i y X_j son independientes (hipótesis nula del test), donde N es el núnero de muestras disponibles (haplotipos de la base de datos), y r_i y r_j son el número de estados de las variables X_i y X_j respectivamente. En nuestro caso concreto, $r_i = r_j = 4$ y por lo tanto $d = 9$.

En nuestro caso para $d = 9$, $N = 1002$, el test rechaza la hipótesis nula (que X_i y X_j son independientes) para una significatividad de $\alpha = 0{,}01$, cuando $I(X_i, X_j) = \frac{t_{\alpha=0,01}}{2N} = \frac{21{,}66}{2\times 1002} = 0{,}01$. Este valor es el límite a partir del cuál la hipótesis nula será rechazada, por lo que ese par de posiciones (Información Mutua) no serán independientes. Este test sera la principal herramienta que emplearemos para analizar las muestras disponibles (haplotipos de la base de datos).

2. Análisis de las muestras

2.1. Objetivo

El objetivo del análisis es verificar si podemos considerar o no las posiciones independientes entre sí. Para ello analizaremos las relaciones (dependencias) entre pares de variables empleando la información mutua. Los haplotipos de las muestras, como vimos anteriormente, constan de aquellas posiciones cuyo valor es diferente al haplotipo de referencia. La idea es ver si conociendo el valor de la posición Y, podemos deducir el valor de la posición X.

2.2. Preprocesado

Los datos disponibles -los haplotipos- están recogidos de un forma poco útil para su tratamiento. Por esta razón, hay que realizar un procesamiento previo de ordenación y filtrado. De esta forma, para cada haplotipo, ordenamos las posiciones de menor a mayor y, mediante una expresión regular (`/^ [0-9]+[A-Z]$/`), aceptamos únicamente aquellas posiciones que cumplan el patrón.

Por ejemplo, si tenemos como entrada la siguiente muestra de un haplotipo,

16311Y,16342C,73G,263G,282C,309.1C,315.1C

Una vez ordenada y filtrada quedaría de la siguiente manera:

73G,263G,282C,16342C

El objetivo de este preprocesado es eliminar las heteroplasmias de posición (representadas por letras distintas de A,C,G,T) y de longitud, representadas con valores de posición que incluyen un punto.

A continuación, generamos una matriz de valores para cada posición. En aquellas posiciones en las que una muestra no tenga valores o sean diferentes de A, C, G o T, establecemos una interrogante (?) si se encuentra fuera del rango del haplotipo o un valor prefijado por nosotros (R) si se encuentra dentro del rango. Este rango indica entre qué posiciones se han analizado en la muestra. De manera que, aquellas posiciones de los rangos medidos que sean diferentes al valor de referencia tendrán su valor (ACTG), aquellas que sean iguales serán R y aquellas posiciones que están fuera de los rangos de la muestra serán interrogante (?), ya que, desconocemos su valor.

Esta información se almacena en forma de vector y tendrá tantas posiciones como posiciones diferentes haya en el conjunto de haplotipos. Es decir, si un haplotipo tiene la posición 75 y otro la posición 340, los vectores de los haplotipos tendrán ambas posiciones.

Siguiendo el ejemplo anterior, supongamos que el rango de dicho haplotipo fuera $16024 - 16383$ y $66 - 370$. En este caso obtendríamos lo mostrado en la Figura 6.

Posición	...	65	66	...	72	73	74	...	262	263	264	...	281	282	283	...	370	371	...	16024	16025	...	16342	...	16383	16384	...
Haplotipo	?	?	R	R	R	G	R	R	R	G	R	R	R	C	R	R	R	?	?	?	R	R	C	R	R	?	?

Figura 6: Representación de un haplotipo.

2.3. Resultados

Una vez disponemos de la matriz con los datos preprocesados, realizamos el cálculo de la información mutua para cada par de posiciones disponibles. Obteniendo, así, una matriz simétrica de informaciónes mútuas donde la diagonal son las entropias, ya que, se está calculando la información mútua de la posición iésima consigo misma.

A fin de analizar visualmente los resultados obtenidos, generamos una imagen en la que la Información mutua se representa por medio de una escala de color (ver Figura 7).

Como se puede observar, la mayor parte de la Figura 7 se encuentra en blanco, salvo algunos puntos que aparecen en una escala de grises [5].

En la esquina inferior izquierda se encuentran las posiciones menores $(1, 2, \dots)$, mientras que van en aumento al ascender hacia la esquina superior derecha $(16500, 16501, \dots)$. Hay que tener en cuenta que estas posiciones en su final y en el comienzo se encuentran enlazadas, de ahí que en la esquina inferior derecha, haya tantas posiciones resaltadas en gris oscuro o negro. Esto es debido a que estas últimas posiciones está muy póximas a las primeras posiciones.

También se puede apreciar que algunas filas poseen varios puntos resaltados en gris formando líneas verticales u horizontales. Estas posiciones se corresponden con la zona situada entre HVR1 y HVR2, la cuál sólo ha sido medida en algunas de las muestras.

En el caso de las rayas horizontales que se observan un poco antes de la parte central de la matriz simétrica, podríamos suponer que todas esas posiciones se encuentran correlacionadas. Aunque la verdad es que, de momento, se desconoce la razón por la que aparece así, puesto que, de estar correlacionadas todas esas posiciones, no aparecerían como líneas, sino como un enorme cuadrado gris y no es lo que

[5] Hay que destacar que con el fin de poder visualizar mejor los datos, se ha aumentado el contraste elevando los valores de la matriz de informaciones mútuas a 0,85.

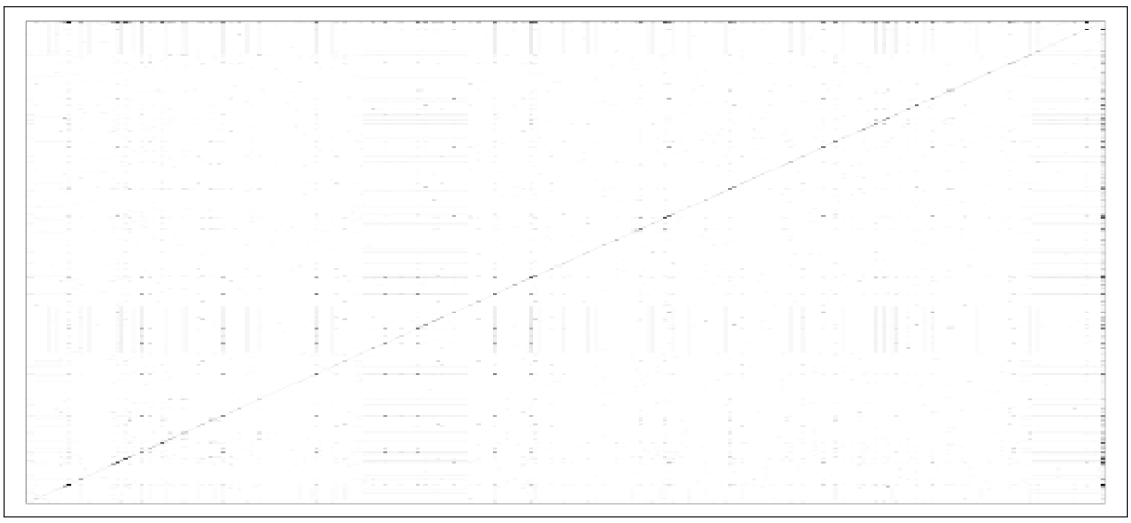

Figura 7: Matriz de Información Mútua de las posiciones de los haplotipos.

sucede.

Con la idea de ver la información que nos brinda la matriz de informaciones mútuas, vamos a transformarla en una matriz binaria determinando que valores están por encima del umbral $I = 0{,}01$, que es el valor a partir del cual el test dice que no son independientes. La matriz resultante se puede ver en la Figura 8.

Lo que observamos en la Figura 8 es que, aunque existe una dependencia entre unas pocas posiciones, en general la reducción de la incertidumbre en una posición conociendo otra no es significativa. Es decir, podemos asumir que las posiciones son independientes unas de otras.

Una conclusión similar se puede extraer de la Figura 10, donde se muestra el histograma con las frecuencias de los valores de la matriz de información mútua. Como se puede ver, casi todos los valores de la matriz se encuentran entre 0 y $0,05$. Unos pocos valores superan el umbral de 0,01, que son los puntos negros que se pueden ver en la Figura 10. Como en el histograma no se puede observar en detalle la distribución de los valores de las informaciones mutuas, se ha obtenido en R (ver Figura 9) el porcentaje de informaciones mutuas que superan el umbral indicado (0,01), a partir del cual las posiciones no son independientes. Como resultado se obtiene que sólo un $3,6\,\%$ de las posiciones no son independientes.

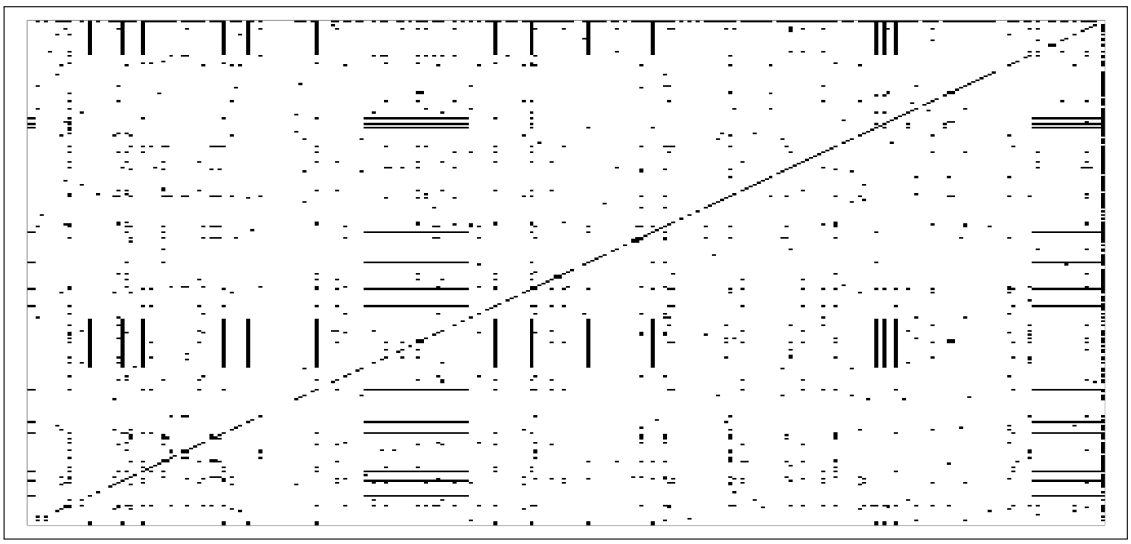

Figura 8: Matriz de informaciones mútuas filtrada a valores mayores de 0,01.

```
> prop.table(table(valores > 0.01))

     FALSE       TRUE
0.96399489 0.03600511
```

Figura 9: Porcentajes de informaciones mutuas que son mayores y menores a 0,01.

3. Cálculo del nivel de confianza en la comparación de haplotipos

Partiendo de la conclusión obtenida en el apartado anterior (independencia de las posiciones), hay que definir una métrica que refleje lo que nos interesa: dados dos haplotipos parcialmente observados, obtener el grado de confianza de que ambos sean el mismo.

Para que dos haplotipos se puedan comparar, las posiciones cuyos valores son diferentes a los de referencia y que están dentro de los rangos analizados deben ser iguales, si no es así, el haplotipo resultado será descartado, puesto que no será igual.

Por ejemplo, supongamos que el usuario ha realizado una búsqueda con el siguiente haplotipo que ha

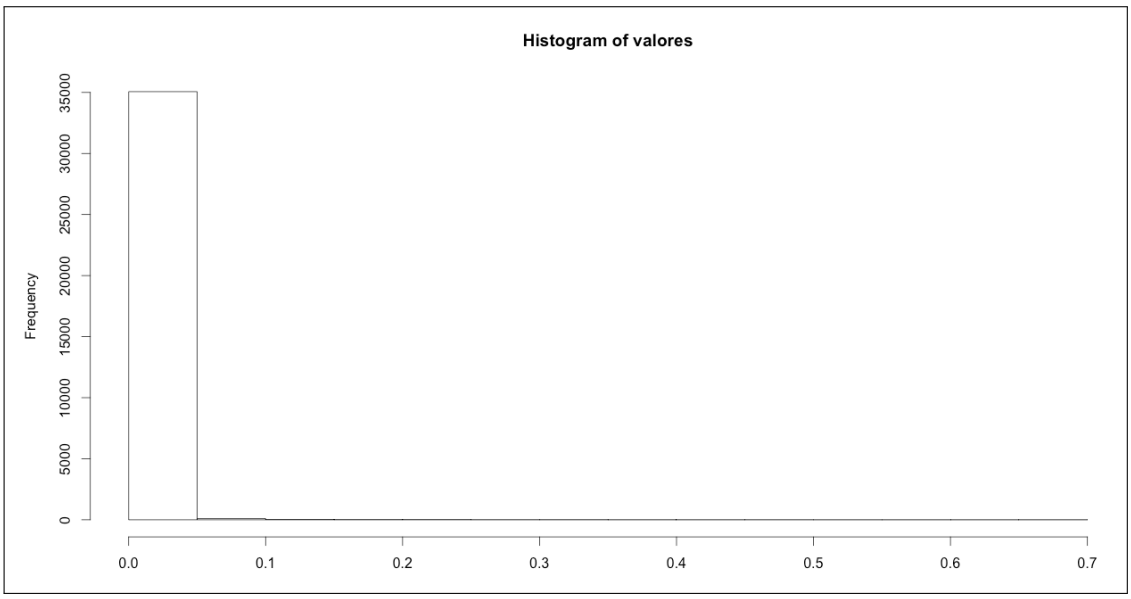

Figura 10: Histograma de los valores de las informaciones mútuas.

sido analizado en los rangos HVR1 y HVR2,

72A,103G,359T,16500G

Al realizar la búsqueda en la base de datos se obtiene como resultado el siguiente haplotipo que también ha sido analizado en HVR1 y HVR2,

72A,103G,359T,16500G

En este caso, al comparar las posiciones y teniendo en cuenta los rangos podemos decir que son iguales. En el caso de que la letra de una sola posición sea diferente entre ellos o que en alguno de ellos haya una más o menos, el haplotipo resultado será descartado y no pasará a la fase de cálculo de confianza. A la hora de calcular la confianza, compararemos los haplotipos posición a posición.

Los casos con los que nos podemos encontrar al comparar dos posiciones serán los siguientes:

- Las dos posiciones se han medido y son iguales. En este caso, la probabilidad de que ambos haplotipos sean el mismo en base a esta posición, será 1.

- Las dos posiciones se han medido y son diferentes. En este caso, es claro que ambos son distintos y, por tanto, la probabilidad será 0.
- Sólo se ha medido la posición en uno de ellos. En este caso, determinaremos la confianza como la probabilidad de observar el valor que sí hemos medido respecto al valor más probable. Es decir, si tenemos dos haplotipos en los que en uno de ellos $X_i = A$ y en el otro $X_i =?$, siendo la distribución X_i $P(X_i = A) = 0{,}2; P(X_i = C) = 0{,}1; P(X_i = G) = 0{,}2; P(X_i = R = T) = 0{,}5$, la confianza será $\frac{P(X_i=A)}{max(P(X_i))} = \frac{P(X_i=A)}{P(X_i=R)} = \frac{0{,}2}{0{,}5} = 0{,}4$.

En general no tenemos una posición sino m y, bajo la suposición de independencia, la confianza total la calculamos como el producto de las probabilidades de la comparación de posiciones.

$$\prod_{i=1}^{m} \frac{P(X_i obs_i)}{max(P(X_i))}$$

Esta ecuación nos dará un índice basado en probabilidades al comparar cada posición de los haplotipos, donde el numerador será la probabilidad de que en la posición iésima haya la letra indicada, y en el divisor mayor probabilidad de aparición de la letra para esa posición. En el caso de que ambas probabilidades sean la misma, el factor máximo será 1.

Por ejemplo, supongamos que nos encontramos con la siguiente posición iésima del segundo rango, donde el haplotipo superior ha sido analizado sólo en un rango (HVR1), mientras que en el haplotipo inferior ha sido analizado en ambos rangos (HVR1 y HVR2). Cuando queremos estudiar las posiciones del segundo rango, no tenemos valores (letras) para las posiciones en HVR2 del haplotipo superior (sólo tenemos posiciones para HVR1), por lo que establecemos interrogantes para ellas. Para las posiciones del haplotipo inferior, que está analizado en HVR1 y HVR2, obtendremos las probabilidades de que aparezcan sus letras. En aquel caso en el que la posición del haplotipo inferior sea la de referencia y el del haplotipo superior sea una interrogante, asignaremos al valor de referencia la probabilidad mayor que haya para esa posición.

Veamos un ejemplo (ver Figura 11). Tenemos dos haplotipos de diez posiciones cada uno. En algunas posiciones ambos haplotipos coinciden en su valores, ya sea porque tienen el valor referencia o porque, aun siendo diferente del valor referencia, son iguales. En estos casos, como hemos visto, la probabilidad de que ambos haplotipos sean iguales es de 1 por posición. Como partimos de la suposición de que son independientes, el producto de todas estas posiciones será 1. Teniendo esto en cuenta, nos podemos

centrar exclusivamente en aquellas posiciones donde sólo se ha medido una de las posiciones.

Supongamos que para las posiciones conflictivas las probabilidades obtenidas para cada valor son las siguientes.

Posición 4	Posición 5	Posición 9
$P_4(A) = 0{,}2$	$P_5(A) = 0{,}6$	$P_9(R = A) = 0{,}2$
$P_4(C) = 0{,}1$	$P_5(R = C) = 0{,}2$	$P_9(C) = 0{,}1$
$P_4(G) = 0{,}2$	$P_5(G) = 0{,}1$	$P_9(G) = 0{,}3$
$P_4(R = T) = 0{,}5$	$P_5(T) = 0{,}1$	$P_9(T) = 0{,}4$

Los resultados que obtenemos son los siguientes:

- Para la posición 4:

$$\frac{P_4(X_4=R=T)}{max(P(X_4))} = \frac{0{,}5}{0{,}5} = 1$$

- Para la posición 5:

$$\frac{P_5(X_5=R=C)}{max(P(X_5))} = \frac{0{,}2}{0{,}6} = 0{,}33$$

- Para la posición 9:

$$\frac{P_9(X_9=R=T)}{max(P(X_9))} = \frac{0{,}2}{0{,}4} = 0{,}5$$

Como hemos dicho, partimos de la suposición de que las posiciones son independientes, por lo que, para obtener la probabilidad de que ambos haplotipos son el mismo, debemos realizar el producto de las probabilidades de todas las posiciones. En este caso, el cálculo del producto de las probabilidades queda de la siguiente manera.

$$\prod_{i=1}^{m} \frac{p_i(obs_i)}{max_i p_i} = 1 \times 1 \times 1 \times 1 \times 0{,}33 \times 1 \times 1 \times 1 \times 0{,}5 \times 1 = 0{,}33 \times 0{,}5 = 0{,}165$$

Obteniendo de esta manera el índice que será utilizado para indicar al usuario el grado de confianza de ese resultado.

POSICIÓN	1	2	3	4	5	6	7	8	9	10
HAPLOTIPO 1	R	R	T	R	A	R	R	R	?	R
HAPLOTIPO 2	R	R	T	?	?	R	R	R	T	R

Figura 11: Ejemplo de cálculo de confianza.

4. Aplicación

4.1. Objetivo

El grupo de investigación Biomicis ha realizado desde su creación diferentes tipos de proyectos relacionados con el ADN mitocondrial, gracias a la generosidad de la gente que ha donado sus análisis clínicos para investigación. Con el tiempo, este grupo de investigación ha querido devolver esta generosidad a la comunidad, ofreciendo una aplicación Web que permita a las personas informarse y, tal vez la parte más interesante, realizar la búsqueda Online del origen de su ADN mitocondrial, a través de su haplotipo.

4.2. Plataforma

Para el desarrollo de la aplicación Web se ha decidido utilizar una plataforma basada en Software Libre/Open Source, que permitirá su creación y posterior mantemiento sin recurrir a licencias de software.

Las tecnologías escogidas son las siguientes:

- Sistema operativo: GNU/Linux
- Servidor Web: Apache 2
- Lenguaje de programación: PHP 5.3. Se trata de un lenguaje ampliamente utilizado en el mundo del desarrollo Web, con abundante información disponible en Internet y orientado a objetos.
- Framework: CodeIgniter. Se escoge este framework, basado en el patrón de arquitectura de software Modelo-Vista-Controlador (MVC), porque permite ahorrar tiempo a la hora de realizar cualquier desarrollo Web y su curva de aprendizaje es inferior a otros framework disponibles como son Symfony

o Zend Framework. Al basarse en el patrón MVC, nos aseguramos de la ?independencia? entre cada una de estas capas, lo que será de mucha utilidad cuando el proyecto vaya haciéndose más grande y gane en complejidad.

  - Gestiona todos los accesos a la información, en nuestro caso a la base de datos. Las peticiones de acceso a la información le llegan a través del controlador.
  - Recibe las peticiones del usuario -eventos- e invoca al modelo cuando se hace alguna solicitud sobre la información. También puede enviar información a la vista asociada según se requiera. Se puede decir que el controlador es un intermediario entre la vista y el modelo.
  - Presenta la información de una forma adecuada al usuario. También es la encargada de generar los eventos que produzca el usuario.

- Otros: Para la interfaz Web se utilizan los frameworks jQuery, como capa de abstracción de Javascript, y para el diseño Bootstrap, lo que permite diseñar de una forma más clara y eficiente un portal Web al contar con un conjunto de librerías de Javascript y hojas de estilo (CSS) ya predifinidas, aunque personalizables. Se utiliza también la API de Google para el uso de mapas incrustados en la aplicación Web, para indicar el origen de los haplotipos obtenidos como resultado de la búsqueda.

4.3. Modelo de datos

El modelo de datos que se utiliza para el almacenaje y búsqueda de información del posible origen mitocondrial de una muestra o haplotipo es bastante sencillo. Este consta de tres tablas:

- Muestra
- Haplotipo
- Haplogrupo

En la tabla Muestra se almacena la información relevante que hace única a una muestra. Cabe destacar que cada muestra tiene la posibilidad de disponer del origen de la muestra, de la madre y/o de la abuela de la muestra, junto con las latitudes y longitudes de la donación y/o de la muestra. La diferencia entre las coordenadas y el origen radica en que, mientras las coordenadas están basadas en latitud y longitud, el origen es un texto descriptivo del lugar, como puede ser ?Donostia (Gipuzkoa)?. Con esta información

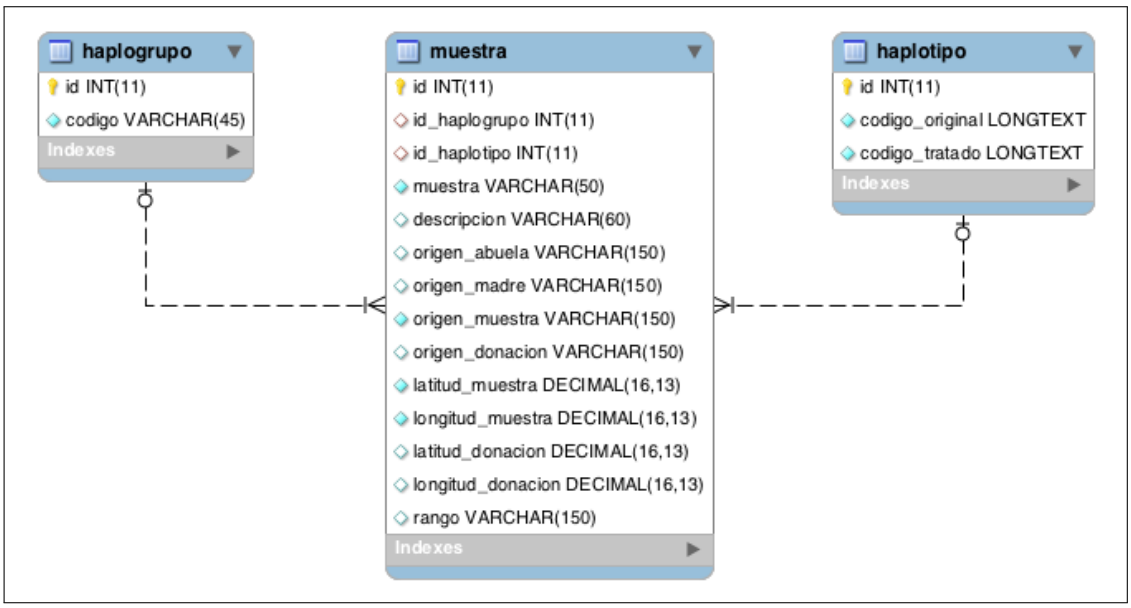

Figura 12: Modelo de datos de la aplicación Web

se pretende poder mostrar en los resultados de las búsquedas la localización greográfica más antigua de que se disponga, con el fin de ver la procedencia de la muestra buscada.

El porqué de que el haplotipo se encuentre en una tabla aparte es debido a que éste puede coincidir entre varias personas, pertenecientes al mismo linaje. De esta manera, si se tiene una gran cantidad de datos se estará consiguiendo evitar la repetición innecesaria de información, a la vez que se posibilitará el añadir información extra del haplotipo. El caso del haplogrupo es el mismo, dado que el haplogrupo puede ser compartido entre varias muestras, se extrae a una tabla con el fin de evitar repetición de información, ahorrar espacio y poder complementar con nueva información en el futuro.

En la siguiente imagen se puede ver los campos que conforman estas tres tablas dentro del modelo de datos.

4.4. Proceso de búsqueda

El proceso de búsqueda consta de las siguientes fases:

- El usuario introduce su haplotipo en la caja de texto y selecciona los rangos adecuados a su análisis

(y pulsa el botón buscar).

- Se realiza una búsqueda en la base de datos para obtener todos aquellos haplotipos que sean susceptibles de coincidir con la muestra del usuario.
- Una vez se tienen todos los resultados, se realiza un filtrado exhaustivo haplotipo a haplotipo mientras se compara con el entregado por el usuario. Si pasa el filtro, se calcula el grado de confianza y se asocia a la información obtenida de la base de datos.
- De los resultados finales se obtienen sus coordenadas y se marcan como chinchetas en el mapa.
- Se visualiza el mapa.
- Se visualiza el listado de resultados.

Cabe destacar que en la versión actual no se ha implementado aún el cálculo de la confianza.

4.5. Imágenes de la aplicación

Lo que a continuación se muestra son algunas capturas de pantalla de la aplicación, con el fin de que se pueda tener una idea mental de su funcionamiento y de la información que, a día de hoy 02 de Septiembre de 2014, se muestra en la Web.

La Figura 13 muestra la pantalla de bienvenida que se muestra nada más acceder al portal Web. Es aquí donde se muestran unos breves mensajes de información sobre la actividad de la Web que rotan cada cierto periodo de tiempo.

En el menú superior se puede ver las distintas opciones de interacción que posee la aplicación Web. La opción más interesante es la que ha llevado al desarrollo de la aplicación, el buscador.

En la sección del buscador se presenta al usuario un formulario, nada complejo, donde deberá introducir su haplotipo e indicar las regiones sobre las que se ha realizado el estudio. A continuación, deberá pulsar el botón buscar y se realizará la búsqueda con su posterior análisis de confianza sobre los resultados obtenidos. Como punto final, se mu estra la tabla con los resultados y, a través de la API de Google, se indica en un mapa de Google Maps las localizaciones más antiguas disponibles para cada uno de los resultados de las muestras de ADNmt (ver Figura 14).

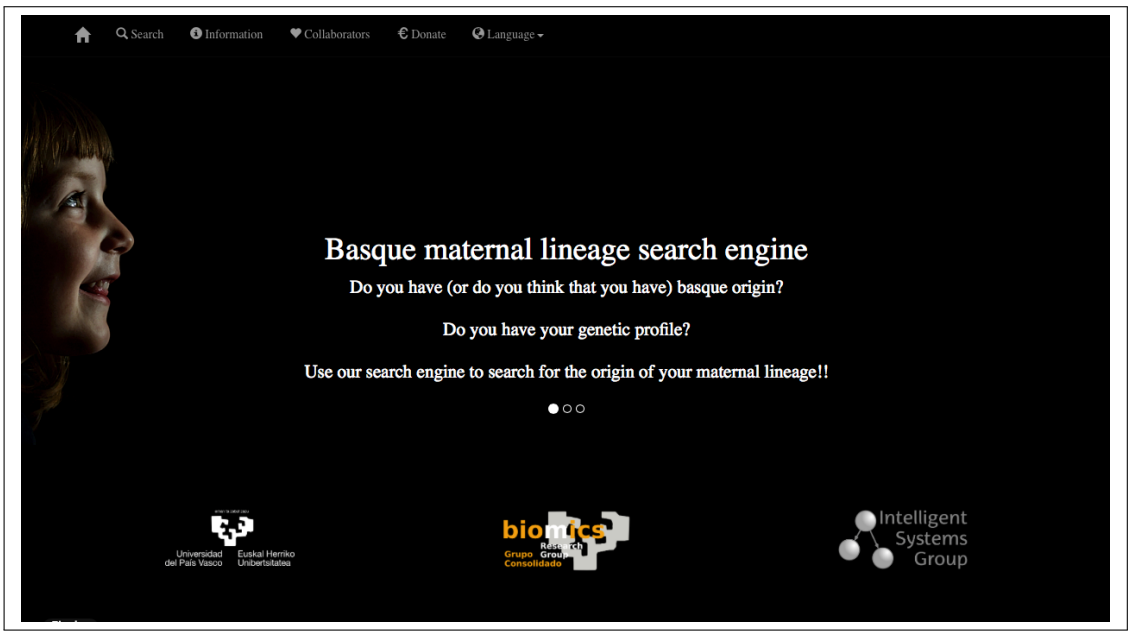

Figura 13: Bienvenida a la aplicación Web.

5. Conclusiones y trabajo futuro

A lo largo de este documento, se ha presentado la necesidad de desarrollar una aplicación que permita comparar haplotipos en presencia de incertidumbre. Para ello se ha utilizado la Teoría de la Información para estudiar la (in)dependencia entre posiciones. El análisis de los datos muestra que salvo algunos casos aislados, las posiciones que conforman los haplotipos son independientes unas de otras.

Llegados a este punto, se muestra y explica la necesidad de una fórmula que permita establecer un grado de confianza en los resultados obtenidos al realizar búsquedas, siendo esta una de las partes más importantes de este documento.

Finalmente, hemos visto características propias del desarrollo de la aplicación, como son las herramientas y tecnologías utilizadas, y algunas capturas de pantalla que nos permiten visualizar el diseño y funcionamiento de la aplicación.

Como trabajo futuro queda pendiente realizar mejorar el cálculo de la confianza incluyendo algunas correlaciones puntuales, vistas en la Figura 7, conocimiento experto relativo a los haplogrupos e incluir

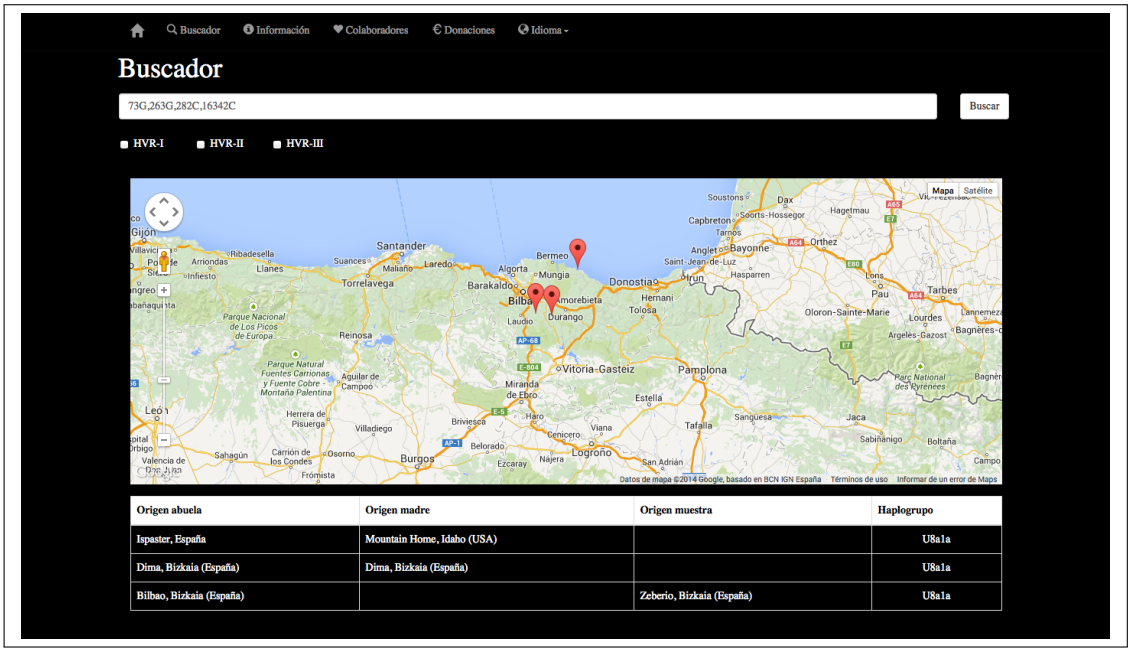

Figura 14: Resultados de la búsqueda de un perfil genético.

la información sobre las heteroplasmias vistas en uno de los ejemplos.

Queda también pendiente implementar en la aplicación Web el cálculo del grado de confianza de los resultados. Para que se muestre así un indicador junto con los resultados, indicando al usuario cuál de los resultados es más cercano a su haplotipo.

Referencias

[1] Cover, M. C. and Thomas, J. A.: Elements of Information Theory. John Wiley and Sons, Inc. (2006)

[2] Kullback, S.: Information Theory and Statistics. Dover Publictions, Inc. (1968)

[3] Andrews: Reanalysis and revision of the Cambridge reference sequence for human mitocondrial ADN. Nature Genetics, 23C2. (1999)